Libros de Cátedra

Probabilidades y Estadística

Análisis de datos

María Carmen Apezteguía y Julieta Ferrario

FACULTAD DE
CIENCIAS EXACTAS

e
exactas





PROBABILIDADES Y ESTADÍSTICA
ANÁLISIS DE DATOS

María Carmen Apezteguía
Julieta Ferrario

(Coordinadoras)

Facultad de Ciencias Exactas



Agradecimientos

Es nuestro deseo agradecer a todos los miembros de la Cátedra Análisis de Datos que formaron parte de este proyecto realizando aportes significativos y críticas constructivas; invirtiendo su tiempo, sus recursos y sus energías para contribuir a la realización del mismo.

Agradecer muy especialmente a la Secretaria de Asuntos Académicos, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, por promover la realización del Libro de Cátedra para la materia Análisis de Datos mediante La Convocatoria 2015 del proyecto que lleva su nombre.

Índice

INTRODUCCIÓN

Este libro intenta dar una introducción a las ideas básicas de Probabilidades y Estadística. Fue concebido para ser usado por los estudiantes del ciclo básico de la Facultad de Ciencias Exactas de la UNLP (CiBEx), con conocimientos básicos de Análisis Matemático.

La temática puede considerarse dividida en dos núcleos centrales:

1. Nociones básicas de probabilidades: son las herramientas necesarias para desarrollar las primeras nociones de inferencia estadística.

2. Algunos elementos de inferencia estadística: estimación puntual y mediante intervalos de confianza, test de hipótesis, regresión lineal.

La gran mayoría de los temas son introducidos con ejemplos, tratando de enfatizar la manera correcta de encararlos.

CAPÍTULO 1

Probabilidades

Introducción: ¿Por qué estudiar Probabilidades?

La Teoría de Probabilidades es una rama de la Matemática, que en sus orígenes se relacionó con la resolución de problemas vinculados con los juegos de azar. Sin embargo, tiene aplicaciones en situaciones muy diversas, ya que se utiliza para estudiar cualquier fenómeno donde no se puede tener certeza del resultado. Este tipo de fenómeno se llama **experimento aleatorio**. Cuando se realizan replicaciones de cualquier medición, por ejemplo en química clínica, se puede observar una variación en los resultados. Esta variación es inherente al proceso de medición. Entonces, el resultado de una medición es incierto, por ese motivo puede considerarse como un experimento aleatorio. La teoría de probabilidades brinda herramientas útiles para manejar este tipo de datos.

Para ejemplificar los primeros conceptos de probabilidad usaremos algunos experimentos aleatorios que se refieren a juegos de azar simples como: arrojar un dado, realizar un tiro de ruleta, sacar una bolilla de una caja con bolillas de diferente color, etc.

Definiciones y propiedades básicas

Espacio muestral. Eventos

Para cada experimento aleatorio existe un conjunto de resultados posibles, llamado **espacio muestral**, denotado por Ω.

Ejemplo 1.1

El lanzamiento de un dado puede dar lugar a 6 resultados: 1, 2, 3, 4, 5, 6, y el espacio muestral o espacio de probabilidad en este caso es $\Omega = \{1, 2, 3, 4, 5, 6\}$.

■

Ejemplo 1.2

El espacio muestral correspondiente al tiro de una ruleta es $\Omega = \{0, 1, 2, \ldots, 36\}$.

■

Ejemplo 1.3

Si se extrae una bolilla de una caja que contiene bolillas rojas, blancas y azules, los posibles resultados son los colores y el espacio muestral correspondiente es $\Omega = \{\text{roja}, \text{blanca}, \text{azul}\}$.

■

Ejemplo 1.4

Si se lanza una moneda tantas veces como sea necesario hasta que sale cara y designamos, por ejemplo, XC al resultado "en el primer lanzamiento sale ceca y en el segundo sale cara", podemos escribir el espacio muestral como $\Omega = \{C, XC, XXC, XXXC, XXXXC, \ldots\}$.

■

Ejemplo 1.5

Si se hace un tiro a un blanco circular de radio r y se determinan las coordenadas del punto de impacto, los resultados posibles son todos los puntos del círculo (para simplificar suponemos el origen de coordenadas en el centro del círculo). En este caso el espacio muestral es

$$\Omega = \{(x, y) \text{ que verifican } x^2 + y^2 \leq r^2\}$$

■



Observación:

En los Ejemplos 1.1, 1.2 y 1.3 el espacio muestral tiene un número finito de elementos: 6, 37 y 3, respectivamente. En el Ejemplo 1.4, el espacio muestral es infinito numerable (sus elementos se pueden enumerar), mientras que en el Ejemplo 1.5, el espacio muestral es infinito no numerable.

EJERCICIO 1.1

Describir en cada una de las siguientes situaciones el espacio muestral, indicando si se trata de un espacio finito, infinito numerable o infinito no numerable.

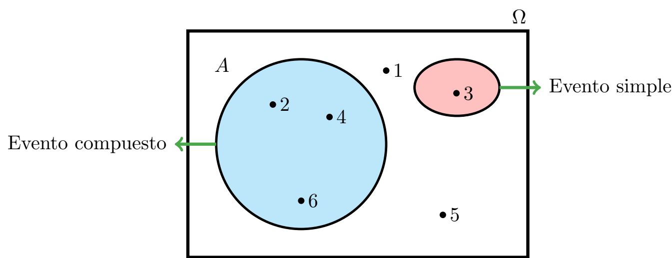
1. Se observa el tiempo en que una máquina trabaja sin romperse.
2. Se quiere contar la cantidad de clavos defectuosos en cajas de 100.
3. Se observa la cantidad de alumnos inscriptos en la materia Análisis de Datos de la Facultad de Ciencias Exactas de la UNLP.

Definición:

A los subconjuntos de Ω se los llama **eventos**. Si un evento está formado por un único resultado será un **evento simple**, en cambio, si consta de más de un resultado, un **evento compuesto**.

Ejemplo 1.6

En el Ejemplo 1.1 los eventos simples son: $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{6\}$; y un ejemplo de evento compuesto es $\{2,4,6\}$. Por convención, a los eventos aleatorios se los suele designar con alguna de las primeras letras del alfabeto en mayúscula, por ejemplo, $A = \{2,4,6\}$. Gráficamente, ésto sería:



Esta grágica se conoce como **Diagrama de Venn**.

■

El espacio muestral es también un evento aleatorio, como sabemos, todo conjunto es subconjunto de sí mismo; también lo es el conjunto vacío $\emptyset$, ya que está incluído en cualquier conjunto, en particular en Ω.

Una vez realizado el experimento, un determinado evento B puede ocurrir o no. Se dice que ocurre cuando el resultado del experimento es un elemento de B, y no ocurre en caso contrario. Como el evento Ω siempre ocurre (por constar de todos los resultados), se dice que es un **evento seguro**; el evento $\emptyset$ que no consta de ningún resultado, como nunca puede suceder, se dice que es un **evento imposible**.

Las operaciones y relaciones habituales entre conjuntos, tienen una traducción intuitiva en términos probabilísticos. Dados dos eventos A y B:

- la intersección, $A \cap B$, es el evento: "A y B ocurren simultáneamente";
- la unión, $A \cup B$, es el evento: "ocurre al menos uno de los dos";
- el complemento de A, A^c, es el evento: "no ocurre A";
- la diferencia, $A - B = A \cap B^c$, es el evento: "ocurre A pero no B";
- si A está incluído en B, $A \subseteq B$, se puede interpretar que: "siempre que ocurre A, ocurre B";
- si A y B no tienen elementos en común, $A \cap B = \emptyset$, entonces: "A y B no pueden ocurrir simultáneamente" y, en este caso, se dice que A y B son **eventos mutuamente excluyentes**, **disjuntos** o **incompatibles**.

En cualquier libro básico de Álgebra se pueden encontrar las nociones fundamentales de Teoría de Conjuntos.

EJERCICIO 1.2

Para resolver los siguientes ejercicios les recomendamos realizar las gráficas de cada uno.

1. Sean Ω el conjunto de los enteros positivos de 1 a 8, $A = \{1, 3, 5\}$, $B = \{1, 4\}$ y $C = \{2, 3, 4, 6\}$. Anote los elementos de los siguientes conjuntos: $A \cap B$, C^c, $\{x \in \Omega : x \in C \text{ y } x \notin B\}$, $(C - B^c) \cap A^c$ y $(A \cup B \cup C)^c$.

2. Sean $\Omega = \{x \in \mathbb{R} : 0 \leq x \leq 2\}$, $A = \{x \in \mathbb{R} : 0.5 < x \leq 1\}$ y $B = \{x \in \mathbb{R} : 0.25 \leq x < 1.5\}$. Describa los siguientes conjuntos: A^c, $(A \cup B)^c$, $A \cup B^c$, $(A \cap B)^c$ y $A^c \cap B$.

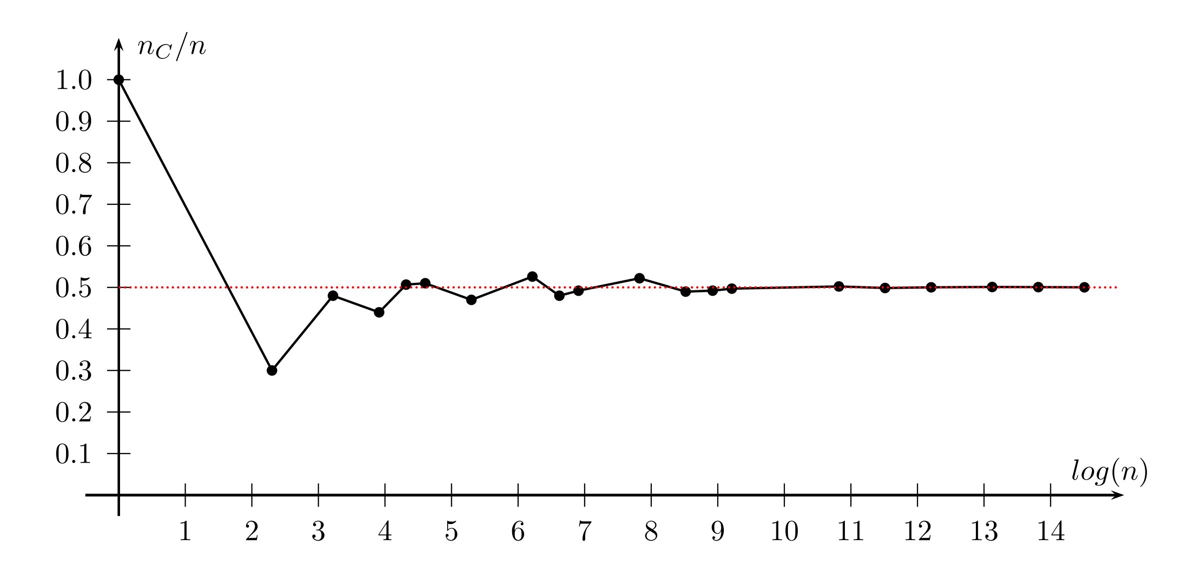
Definición de Probabilidad. Axiomas de Probabilidad

Si bien ante un experimento aleatorio no se puede saber de antemano qué resultado se va a obtener, nos interesa disponer de algún número que mida la posibilidad de que ocurra cada uno de los eventos. Si se lanza una moneda al aire n veces, una medida de la posibilidad de que salga cara (evento C) podría ser la frecuencia relativa de caras, es decir, el número $\text{fr}_C = n_C/n$ (donde n_C indica la cantidad de veces que se obtuvo cara, es decir, el número de ocurrencia del suceso C en

los n lanzamientos). Sin embargo, al lanzar 10 veces la moneda, podría obtenerse una frecuencia relativa de 0.6, en tanto que al lanzarla otras 10 veces distintas, podría conseguirse un valor de 0.4 o cualquier otro número; y no resultaría útil que la medida de la posibilidad de un evento dependa de una experiencia particular, esta medida debe ser un número objetivo. En la siguiente tabla se muestran las frecuencias relativas de caras, n_C/n, en una realización de este experimento:

$\boldsymbol{n}$	10	25	50	75	100	200	500	750	1000
$\boldsymbol{n_C/n}$	0.3000	0.4800	0.4400	0.5067	0.5100	0.4700	0.5260	0.4800	0.4920

Se puede observar que cuando una moneda normal se lanza un número de veces cada vez mayor, la frecuencia relativa de caras se va estabilizando alrededor de un número fijo, 0.5. La siguiente gráfica muestra esta estabilidad, pero para apreciarla se han graficado los puntos $\big(log(n), n_C/n\big)$ en lugar de $(n, n_C/n)$.



La estabilización de las frecuencias relativas de un evento alrededor de un número, ocurre para cualquier experimento aleatorio que se repita muchas veces. Una idea intuitiva de la probabilidad de un evento A, sería el límite de las frecuencias relativas, cuando n tiende a infinito.

Se puede verificar fácilmente que la frecuencia relativa tiene las siguientes propiedades:

- $0 \leq \text{fr}_A = n_A/n \leq 1$ para todo evento A.

- $\text{fr}_\Omega = n_\Omega/n = 1$ (donde Ω es el espacio muestral).

- **Ley aditiva:** Si los eventos A y B son disjuntos:

$$\text{fr}_{A\cup B} = n_{A\cup B}/n = n_A/n + n_B/n = \text{fr}_A + \text{fr}_B$$

Entonces, el límite de esas frecuencias heredaría esas propiedades.

Para que el concepto de probabilidad coincida con esta idea intuitiva, vamos a definirlo de modo que cumpla esas mismas propiedades.

Definición:

Dado un experimento aleatorio con espacio muestral Ω, una **probabilidad** es una función P, que a cada evento A de Ω le asigna un número, llamado probabilidad de A, que se denota $P(A)$, y que verifica:

(A1) $0 \leq P(A) \leq 1$ para todo evento A.

(A2) $P(\Omega) = 1$.

(A3) **Ley aditiva:** Si los eventos A y B son disjuntos, es decir, $A \cap B = \emptyset$,

$$P(A \cup B) = P(A) + P(B).$$

(A4) Si A_1, A_2, A_3, $\ldots$ es una colección infinita de eventos mutuamente excluyentes, es decir, $A_i \cap A_j = \emptyset$, para $i \neq j$, entonces:

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = \sum_{i=1}^{\infty} P(A_i)$$

Aclaración

A partir del **(A3)**, se puede generalizar la propiedad aditiva para n eventos mutuamente excluyentes, pero no puede generalizarse para una colección infinita numerable de eventos, por eso, para trabajar con espacios muestrales infinitos, es necesario agregar el **(A4)**.

A partir de esta definición de probabilidad, pueden deducirse varias propiedades de manera bastante simple.

Algunas propiedades básicas

PROPIEDAD 1.1: Para cualquier evento A, $P(A) = 1 - P(A^c)$.

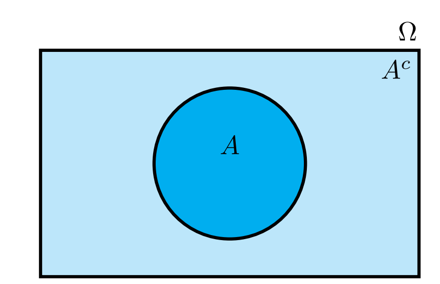
Demostración: Por definición de A^c:

$$A \cup A^c = \Omega \text{ y } A \cap A^c = \emptyset$$

Por **(A2)** y **(A3)**

$$1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$$

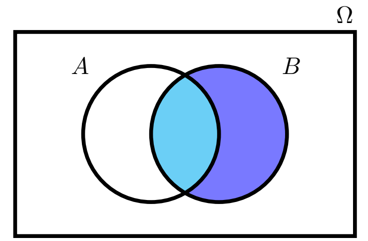
Despejando $P(A)$: $P(A) = 1 - P(A^c)$.



En particular: si $A = \emptyset$, se cumple que $P(\emptyset) = 0$, ya que $A^c = \Omega$. Es importante notar que el recíproco no es verdadero. Si $P(A) = 0$ no se puede concluir que $A = \emptyset$, veremos más adelante que hay eventos no vacíos que pueden tener probabilidad cero.

PROPIEDAD 1.2: Para dos eventos cualesquiera A y B, $P(B - A) = P(B) - P(A \cap B)$. En particular, si $A \subseteq B$, $P(B - A) = P(B) - P(A)$ y $P(B) \geq P(A)$.

Demostración: El evento B puede escribirse como:



$$B = (A \cap B) \cup (B \cap A^c)$$

donde $(A \cap B) =$ y $(B \cap A^c) =$ son disjuntos. Luego, por **(A3)**:

$$P(B) = P(A \cap B) + P(B \cap A^c).$$

Despejando $P(B \cap A^c) = P(B) - P(A \cap B)$. Por último, por la Propiedad del complemento en el Apéndice A, $B - A = B \cap A^c$, por lo tanto $P(B - A) = P(B) - P(A \cap B)$.

En el caso que A esté contenido en B, $A \subseteq B$, tenemos que $A \cap B = A$, entonces

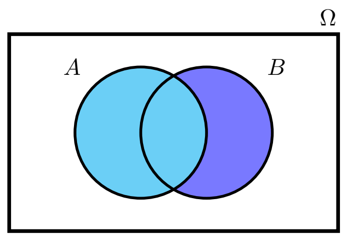
$$P(B - A) = P(B) - P(A), \tag{1.1}$$

por el resultado anterior. Despejando de (1.1) tenemos que $P(B) = P(A) + P(B - A)$. Luego, como $P(B - A) \geq 0$ por **(A1)**: $P(B) = P(A) + P(B - A) \geq P(A)$.

PROPIEDAD 1.3: Para dos eventos cualesquiera A y B,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Demostración: Primero observemos que $A \cup B = A \cup (B \cap A^c)$, donde $A =$ y $(B \cap A^c) =$ son disjuntos.



Por **(A3)**:

$$P(A \cup B) = P[A \cup (B \cap A^c)] = P(A) + P(B \cap A^c) \tag{1.2}$$

Recordemos que, en la Propiedad 1.2, se llegó al siguiente resultado:

$$P(B \cap A^c) = P(B) - P(A \cap B)$$

Reemplazando ésto en (1.2), se obtiene:

$$P(A \cup B) = P(A) + P(B \cap A^c) = P(A) + P(B) - P(A \cap B)$$

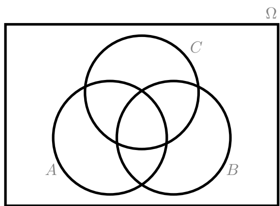
Notar que si $A \cap B = \emptyset$, en esta última propiedad, obtenemos **(A3)**.

Aclaración

Para tres eventos cualesquiera A, B y C, la probabilidad de la unión es:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Para demostrarlo formalmente, se puede escribir $A \cup B \cup C$ como $(A \cup B) \cup C$ y aplicar la Propiedad 1.3.



EJERCICIO 1.3

En los siguientes ejercicios aplicar las propiedades anteriores.

1. Sean A y B dos eventos disjuntos tales que $P(B) = 0.72$ y $P(A \cap B^c) = 0.02$. Calcular $P(A)$ y $P(A \cup B)$.

2. Determinar si las siguientes afirmaciones son falsas o verdaderas. Justificar su respuesta.

   - Si $P(A) > 0$, entonces $P(A \cup B) > 0$.
   - Si $P(A) > 1/2$ y $P(B) > 1/2$, entonces $P(A \cap B) > 0$.
   - Si $P(A) > 0$, entonces $P(A^c) > 0$.
   - $P(B \cup A) = P(B) + P(A)$.

Determinación de probabilidades en espacios muestrales finito o infinito numerables

Cuando el espacio muestral es finito o infinito numerable, para definir una probabilidad sobre todos los eventos, es suficiente asignar probabilidades $P(E_i)$ para todos los eventos simples E_i. Esta asignación debe satisfacer:

- $P(E_i) \geq 0$
- $\sum_i P(E_i) = 1$

Entonces, por **(A3)**, la probabilidad de cualquier evento compuesto A se calcula sumando las $P(E_i)$ para todos los E_i en A

$$P(A) = \sum_{E_i \subseteq A} P(E_i)$$

Ejemplo 1.7

Consideremos el experimento que consiste en tirar un dado que no está bien equilibrado, y resulta que cualquiera de los resultados pares tiene el doble de probabilidad de ocurrir que cualquiera de los resultados impares.

Llamamos E_1, E_2, E_3, E_4, E_5 y E_6 a los eventos simples $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$ y $\{6\}$, respectivamente. La única asignación de probabilidades posible deberá cumplir:

- $P(E_1) = P(E_3) = P(E_5) = 1/9$
- $P(E_2) = P(E_4) = P(E_6) = 2/9$

Luego, la probabilidad de cualquier evento se calcula a partir de esos eventos simples. Por ejemplo, para el evento

$$A = \text{ "el resultado es par" } = \{2,4,6\} = \{2\} \cup \{4\} \cup \{6\} = E_2 \cup E_4 \cup E_6,$$

luego

$$P(A) = P(E_2 \cup E_4 \cup E_6) = P(E_2) + P(E_4) + P(E_6) = 6/9 = 2/3$$

Para

$$B = \text{"el resultado es menor o igual a 3"} = \{1,2,3\} = \{1\} \cup \{2\} \cup \{3\} = E_1 \cup E_2 \cup E_3,$$

entonces

$$P(B) = P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) = 1/9 + 2/9 + 1/9 = 4/9$$

■

EJERCICIO 1.4

Resolver los siguientes ejercicios:

1. Sean A y B dos eventos tales que: $P(A) = 0.2$, $P(B) = 0.3$ y $P(A \cap B) = 0.1$. Calcular:
   - $P(A \cup B)$
   - $P(A^c \cup B^c)$
   - $P(A \cap B^c)$
   - $P(A^c \cup B)$
2. Se construye un dado de manera que el 1 y el 2 ocurran con el doble de frecuencia que se presenta el 5, el cual ocurre con la frecuencia 3 veces superior al 3, al 4 o al 6. Si se lanza una vez, ¿cuál es la probabilidad de que el número sea par? y ¿cuál es la probabilidad de que el número sea mayor que 4?

Espacios equiprobables

Definición:

Un espacio muestral finito, se dice **equiprobable**, si todos los eventos simples, E_i, tienen la misma probabilidad. En ese caso, para que se cumplan las condiciones:

- $P(E_i) \geq 0$
- $\sum_{i=1}^{n} P(E_i) = 1$

la única posible asignación de probabilidades debe ser: $P(E_i) = 1/n$, donde n es el número de elementos del espacio muestral.

Entonces, si A es un evento que está formado por k eventos simples,

$$P(A) = \sum_{E_i \subseteq A} P(E_i) = \frac{k}{n}$$

En consecuencia, en un espacio muestral finito equiprobable, la probabilidad de un evento se calcula como el número de resultados que forman ese evento dividido por el número de resultados de todo el espacio muestral:

$$P(A) = \frac{\# A}{\# \Omega}$$

Ejemplo 1.8

Consideremos el experimento que consiste en tirar un dado equilibrado, en este caso $\# \Omega = 6$ y los 6 resultados tienen igual probabilidad, 1/6.

Sea A = "sale un número par" = $\{2, 4, 6\}$, entonces

$$P(A) = \frac{\# A}{\# \Omega} = 3/6$$

Sea B = "sale un número menor que 5" = $\{1, 2, 3, 4\}$, entonces

$$P(B) = \frac{\# B}{\# \Omega} = 4/6$$

■

Ejemplo 1.9

Consideremos el experimento que consiste en arrojar dos veces un dado equilibrado, para este experimento podemos escribir el espacio muestral como:

$$\begin{aligned}\Omega &= \{(x, y) : \text{donde } x \text{ e } y \in \{1, 2, \ldots, 6\}\} \\ &= \{(1,1); (1,2); \ldots; (1,6); (2,1); (2,2); \ldots; (2,6); \ldots; (6,1); (6,2); \ldots; (6,6)\}\end{aligned}$$

Este espacio muestral es equiprobable y tiene 36 eventos simples, cada uno con probabilidad $1/36$. Sea el evento $A =$ "la suma de los dos resultados es menor que 6",

$$A = \{(1,1); (1,2); (2,1); (1,3); (3,1); (1,4); (4,1); (2,2); (2,3); (3,2)\}$$

entonces $P(A) = 10/36$.

■

Ejemplo 1.10

Consideremos el experimento aleatorio que consiste en sacar una bolilla de una caja que contiene 4 bolillas blancas, 4 rojas y 2 azules. Podemos pensar el espacio muestral formado por todas las extracciones posibles que son 10, y todas tienen igual probabilidad $1/10$.

Luego si definimos el evento $B =$ "sale una bolilla blanca", la $P(B) = 4/10 = 0.4$.

En general si en la caja hay un 40 % de bolillas blancas, $P(B) = 0.4$.

■

Ejemplo 1.11

Consideremos que deseamos calcular la probabilidad de que un individuo elegido en una población tenga determinada característica. Supongamos que se conoce que el 46 % de los individuos de una población tienen sangre del grupo O, el 43 % del grupo A, el 8 % del grupo B y el 3 % del grupo AB. Se elige una persona al azar en dicha población, esto significa que todos los individuos tienen la misma probabildad de ser elegido. Como sabemos que el 46 % de los individuos tiene grupo O, la probabilidad de que el individuo elegido tengo grupo O es 0.46. Del mismo modo la probabilidad de que tenga sangre grupo A es 0.43 y la probabilidad de que tenga sangre grupo A o grupo B es 0.51 (dado que tener sangre grupo A o tener sangre grupo B son eventos incompatibles o disjuntos).

■

EJERCICIO 1.5

En una repisa hay 10 libros distintos de novelas y 20 libros distintos de cuentos, de los cuales la mitad de las novelas y la mitad de los cuentos están escritos en español. Si se selecciona uno de estos libros al azar, hallar la probabilidad de que:

- el libro seleccionado sea una novela;
- el libro seleccionado este escrito en español;
- el libro seleccionado sea una novela y este escrito en español;
- el libro seleccionado sea una novela o este escrito en español.

Probabilidad condicional

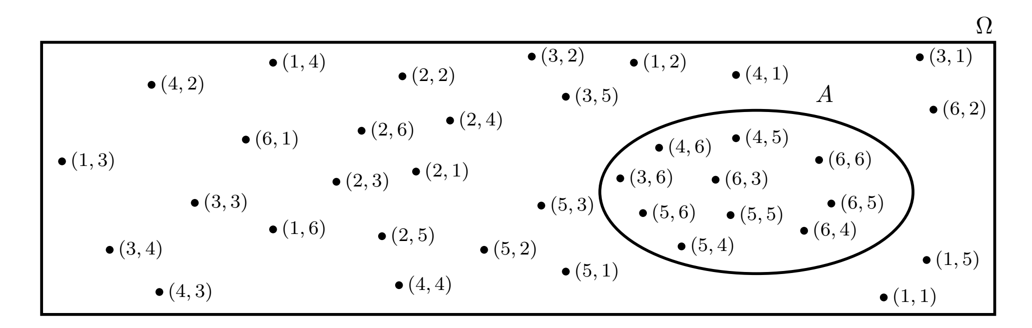
Consideremos el siguiente ejemplo: se arroja dos veces un dado, entonces el espacio muestral se puede definir como

$$\Omega = \{(i,j) : i \text{ es el número del primer tiro y } j \text{ el número del segundo tiro, con } i,j = 1,2,3,4,5,6\}.$$

Nos interesa calcular la probabilidad del evento

$$A = \text{“la suma de los dos resultados es mayor que 8”} = \{(3,6);(4,5);(4,6);(5,4);(5,5);(5,6);(6,3);(6,4);(6,5);(6,6)\}$$

Si el dado es equilibrado $P(A) = 10/36$.



Ahora supongamos que sabemos que en el primer tiro salió un 2, es decir, ocurrió el evento

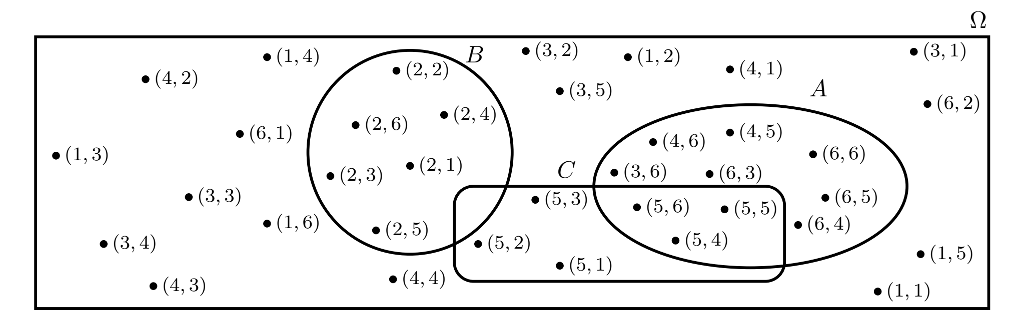
$$B = \text{“el primer tiro es 2”} = \{(2,1);(2,2);(2,3);(2,4);(2,5);(2,6)\}.$$

y en ninguno de estos posibles resultados la suma puede ser mayor que 8. Con esta información es imposible que la suma sea mayor que 8. Usamos la notación $P(A|B)$ para indicar la probabilidad de que ocurra A, sabiendo que ocurrió B. Entonces, en este caso, $P(A|B) = 0$.

Por otra parte, si sabemos que en el primer tiro salió 5, o sea ocurrió

$$C = \{(5,1);(5,2);(5,3);(5,4);\ (5,5);(5,6)\}$$

y sólo en tres de ellos se cumple que la suma es mayor que 8, entonces $P(A|C) = 3/6$.



Consideremos otro ejemplo, se selecciona al azar un recién nacido y se realiza un análisis para diagnosticar hipotiroidismo congénito (HC). Sea $A =$ “el recién nacido padece HC”, la $P(A)$ es igual a la proporción de recién nacidos con HC en la población. Ahora bien, si observamos que el recién nacido es una niña (sea $B =$ “el recién nacido es de sexo femenino”) y queremos conocer la

probabilidad de que padezca HC, esto es la proporción de recién nacidos con HC en esa subpoblación (recién nacidos de sexo femenino). En este ejemplo $P(A|B) > P(A)$, pues es sabido que el HC es más frecuente en las niñas.

Dado que ocurrió B, el espacio muestral pertinente ya no es Ω sino que consiste en los resultados de B. En este caso, A ocurre si y sólo si ocurre uno de los resultados de la intersección $A \cap B$, así que la probabilidad condicional de A dado B es proporcional a $P(A \cap B)$.

Definición:

Dados dos evento A y B, si $P(B) > 0$ se define $P(A|B)$ como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Obviamente, si $P(A) > 0$, también puede definirse

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

Se puede probar que fijando el evento condicionante B, la probabilidad condicional dado B, cumple los axiomas de probabilidad:

(A1) $0 \leq P(A|B) \leq 1$ para cualquier A

(A2) $P(\Omega|B) = 1$ (donde Ω es el espacio muestral)

(A3) **Ley aditiva:** Si los eventos A y C son disjuntos:

$$P(A \cup C|B) = P(A|B) + P(C|B).$$

(A4) Si A_1, A_2, A_3, ... es una colección infinita de eventos mutuamente excluyentes, entonces

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots |B) = \sum_{i=1}^{\infty} P(A_i|B).$$

Por lo tanto, tiene todas las propiedades de una probabilidad.

Ejemplo 1.12

Supongamos que en la población general hay 49 % de hombres y 51 % de mujeres, y que la proporción de hombres y mujeres daltónicos se muestra en la siguiente tabla de probabilidad:

Datos	Hombres	Mujeres	Total
Daltónicos	0.038	0.002	0.040
No daltónicos	0.452	0.508	0.960
Total	0.490	0.510	1

Si se escoge al azar una persona de esta población y se encuentra que es hombre (evento B = "la persona seleccionada es hombre"), ¿cuál es la probabilidad de que sea daltónica (evento A = "la persona seleccionada es daltónica")?

Sabiendo que B ha ocurrido, debemos restringir nuestra atención a sólo 49 % de la población que es de hombres. La probabilidad de ser daltónico, dado que la persona es hombre, es:

$$P(A|B) = P(A \cap B)/P(B) = 0.038/0.49 = 0.078$$

Significa que si sabemos que la persona seleccionada es hombre, este hecho aumenta la probabilidad de que sea daltónico que era 0.04.

Ahora nos preguntamos, ¿cuál es la probabilidad de ser daltónico, dado que la persona es mujer? En este caso estamos restringiendo a sólo el 51 % de la población que es de mujeres y por lo tanto:

$$P(A|B^c) = P(A \cap B^c)/P(B^c) = 0.002/0.51 = 0.004$$

Podemos deducir de este cálculo que la información adicional de que la persona seleccionada es mujer, disminuye la probabilidad de que sea daltónica.

■

EJERCICIO 1.6

Determinar si las siguientes afirmaciones son falsas o verdaderas. Justificar su respuesta.

1. $P(A|B) + P(A^c|B) = 1$.
2. $P(A|B) + P(A|B^c) = P(A)$.
3. $P(A|A \cap B) = P(B|A \cap B) = 1$.
4. $P(A|A) = P(A)$.
5. Si B y C son eventos disjuntos, $P(A|B \cup C) = P(A|B) + P(A|C)$.
6. Si A y B son eventos disjuntos, $P(A \cup B|C) = P(A|C) + P(B|C)$.

Regla de la multiplicación

A partir de la definición de probabilidad condicional podemos deducir dos ecuaciones:

- si $P(B) > 0$, $P(A|B) = P(A \cap B)/P(B)$ implica que $P(A \cap B) = P(A|B) \times P(B)$,
- si $P(A) > 0$, $P(B|A) = P(A \cap B)/P(A)$ implica que $P(A \cap B) = P(B|A) \times P(A)$.

Formalizando:

REGLA DE LA MULTIPLICACIÓN: Dados dos eventos A y B la probabilidad de la intersección puede calcularse como:

$$P(A \cap B) = P(A|B) \times P(B) = P(B|A) \times P(A),$$

cuando estén definidas las respectivas probabilidades condicionales.

La extensión de la regla anterior a tres eventos es:

$$P(A \cap B \cap C) = P(C|A \cap B) \times P(B|A) \times P(A),$$

y de modo similar para más de tres.

Ejemplo 1.13

En un banco de sangre, 4 individuos han respondido a una solicitud. Se necesita sangre tipo A^+ y sólo uno de ellos tiene ese tipo, pero no se sabe cuál. Si los donantes potenciales se seleccionan al azar para determinar su tipo sanguíneo, ¿cuál es la probabilidad de que haya que determinar el tipo sanguíneo en al menos tres individuos para obtener el tipo deseado?

Llamemos B = "primer donante no es A^+" y A = "segundo donante no es A^+", sabemos que $P(B) = 3/4$ y $P(A|B) = 2/3$. El evento $A \cap B$ es:

$$\begin{aligned} A \cap B &= \text{"ni el primero ni el segundo son tipo } A^+\text{"} \\ &= \text{"se determina el tipo sanguíneo en al menos tres individuos".} \end{aligned}$$

Usando la Regla de la multiplicación:

$$P(A \cap B) = P(A|B) \times P(B) = 2/3 \times 3/4 = 1/2$$

■

Eventos independientes

Si volvemos al Ejemplo 1.12, del daltonismo y el género, hemos visto que $P(A|B) \neq P(A)$, con lo cual la probabilidad de que la persona elegida al azar sea daltónica sabiendo que es hombre es distinto a la probabilidad de que esa persona sea daltónica sin saber su género. Es decir, saber que la persona elegida al azar es hombre modifica la probabilidad de que sea daltónica. Eso indicaría que hay alguna relación o dependencia entre los eventos B y A.

Pensemos en otro ejemplo.

Ejemplo 1.14

Se tira un solo dado dos veces y los eventos de interés son: $A =$ "se observa un 2 en el primer tiro" y $B =$ "se observa un 2 en el segundo tiro". Si el dado no está cargado, la probabilidad del evento A es $1/6$, y es lógico pensar que la probabilidad de B también es $1/6$ sin importar si en el primer tiro ocurrió A o no, es decir $P(B) = P(B|A) = P(B|A^c)$, eso significa que los eventos A y B no están relacionados o que son "independientes".

■

Daremos una definición de independencia ligeramente distinta.

Definición:

Los eventos A y B son **independientes** si y sólo si $P(A \cap B) = P(A) \times P(B)$.

A partir de esta definición de independencia, se puede ver que si A y B son eventos independientes y $P(B) > 0$, se cumple $P(A|B) = P(A)$.

La demostración es elemental, ya que

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \times \cancel{P(B)}}{\cancel{P(B)}} = P(A).$$

Entonces la definición de independencia coincide con la idea intuitiva de que saber que ocurrió B, no modifica la probabilidad de que ocurra A.

Definición:

Decimos que los tres eventos A, B y C son **mutuamente independientes** si y sólo si todas las condiciones siguientes se mantienen:

$$P(A \cap B) = P(A) \times P(B)$$
$$P(A \cap C) = P(A) \times P(C)$$
$$P(B \cap C) = P(B) \times P(C)$$
$$P(A \cap B \cap C) = P(A) \times P(B) \times P(C)$$

Definición:

Los n eventos A_1, A_2, $\ldots$, A_n son **mutuamente independientes** si para todo k $(k = 2, 3, \ldots, n)$ y todo subconjunto de índices i_1, i_2, $\ldots$, i_k, se cumple:

$$P\big(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}\big) = P\big(A_{i_1}\big) \times P\big(A_{i_2}\big) \times \cdots \times P\big(A_{i_k}\big)$$

Observación:



La definición de independencia indica que si queremos verificar si dos eventos son independientes, debemos ver que la probabilidad de la intersección es el producto de las probabilidades. Sin embargo, cuando por la naturaleza del experimento aleatorio sabemos que hay independencia, como los dos tiros de un dado, esta definición nos permite calcular la probabilidad de la intersección como el producto de las probabilidades.

En el Ejemplo 1.14, del dado, podemos calcular $P(A \cap B) = 1/6 \times 1/6$.

Por supuesto, si dos eventos no son independientes, la probabilidad de que ocurran simultáneamente no es el producto. Por ejemplo, si la probabilidad de que un hombre tenga una altura superior a 1.80 m es 0.2, la probabilidad de que un padre y un hijo tengan altura superior a 1.80 m no es 0.2×0.2, ya que estos eventos no son independientes (sabemos que la altura de los hijos están relacionadas con la altura de los padres).

PROPOSICIÓN 1.1: Dados dos eventos A y B, las siguientes afirmaciones son equivalentes:

- A y B son independientes.
- A y B^c son independientes.
- A^c y B son independientes.
- A^c y B^c son independientes.

Demostración: Comenzamos probando que la independencia de A y B implica la de A y B^c.
Recordar primero que $A = (A \cap B) \cup (A \cap B^c)$, con ambos conjuntos disjuntos. Luego, aplicando probabilidad, nos queda $P(A) = P(A \cap B) + P(A \cap B^c)$, por **(A3)**.
Ahora, despejando y sabiendo que A y B son independientes, tenemos

$$\begin{aligned}
P(A \cap B^c) &= P(A) - P(A \cap B) && \text{(utilizando la Propiedad 1.2)} \\
&= P(A) - P(A) \times P(B) && \text{(por hipótesis)} \\
&= P(A) \times [1 - P(B)] && \text{(sacando factor común)} \\
&= P(A) \times P(B^c) && \text{(utilizando la Propiedad 1.1)}
\end{aligned}$$

es decir, A y B^c son independientes.
Aplicando este razonamiento a los eventos A y B^c, resulta que la independencia de A y B^c implica la de A y $(B^c)^c = B$, lo que prueba la implicación opuesta. En consecuencia, hemos demostrado que son equivalentes: A y B son independientes y A y B^c son independientes.
De la primera equivalencia salen las otras dos.

EJERCICIO 1.7

1. Sean A y B eventos independientes tales que $P(A) = 0.3$ y $P(B) = 0.24$. Calcular:
   - $P(A \cap B)$
   - $P(A \cup B)$
   - $P(A \cup B^c)$
   - $P(A|B^c)$

2. Sea $\Omega = \{1, 2, 3, 4\}$ un espacio muestral equiprobable. Dados los eventos:

$$A = \{1, 2\}, \quad B = \{2, 3\} \quad \text{y} \quad C = \{2, 4\}.$$

   ¿Son A, B y C independientes?

3. Sean A y B eventos independientes. Demostrar que $P(A \cup B) = 1 - P(A^c) \times P(B^c)$.

Teorema de la Probabilidad Total. Teorema de Bayes

Definición:

Los eventos $A_1, A_2, \ldots, A_n$ representan una **partición del espacio muestral** Ω, si cumplen:

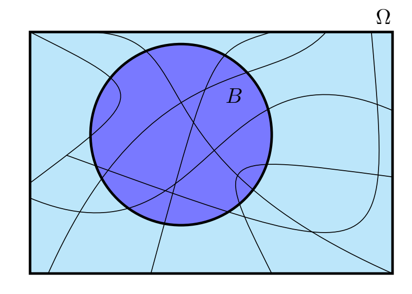
(a) $A_1 \cup A_2 \cup \ldots \cup A_n = \bigcup_{i=1}^{n} A_i = \Omega$ y

(b) $A_i \cap A_j = \emptyset$ para todo $i \neq j$.

TEOREMA DE LA PROBABILIDAD TOTAL: Si $A_1, A_2, \ldots, A_n$ representan una partición del espacio muestral Ω. Y además, $P(A_i) \neq 0$ para todo i. Entonces, para cualquier evento B, se cumple:

$$\begin{aligned} P(B) &= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_n)P(A_n) \\ &= \sum_{i=1}^{n} P(B|A_i)P(A_i). \end{aligned}$$

Demostración: Como los A_i constituyen una partición del espacio Ω, (por **(a)**), cualquier evento B puede escribirse como:



$$
\begin{aligned}
B &= B \cap \Omega \\
&= B \cap \big(A_1 \cup A_2 \cup \ldots \cup A_n\big)
\end{aligned}
$$

Utilizando la propiedad distributiva de la intersección respecto de la unión, tenemos que:

$$B \cap \big(A_1 \cup A_2 \cup \ldots \cup A_n\big) = (B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_n),$$

por lo tanto:

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_n).$$

Como los eventos $(B \cap A_1), (B \cap A_2), \ldots, (B \cap A_n)$ son mutuamente excluyentes (por **(b)**), podemos aplicar la Ley aditiva y escribir:

$$
\begin{aligned}
P(B) &= P\big((B \cap A_1) \cup (B \cap A_2) \cup \ldots \cup (B \cap A_n)\big) \\
&= P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_n).
\end{aligned}
$$

Por la Regla de la multiplicación, cada término $P(B \cap A_i) = P(B|A_i)P(A_i)$ y reemplazando, obtenemos:

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_n)P(A_n).$$

Ejemplo 1.15

En cierta comunidad, el 8 % de los adultos de más de 50 años de edad padece diabetes. Se conoce que la prueba para diagnosticar esa enfermedad tiene una sensibilidad del 95 % (esto significa que si la prueba se aplica a un individuo enfermo, la probabilidad de un resultado positivo es 0.95) y la especificidad es del 98 % (la probabilidad de obtener un resultado negativo dado que el individuo es sano es 0.98).

Recordemos que la prevalencia de una enfermedad en una población se define como la proporción de enfermos en la población, y suele expresarse como porcentaje. En consecuencia, si se elige una persona al azar, la probabilidad de que esté enferma es igual a la prevalencia.

Supongamos que se va a utilizar esta prueba diagnóstica en un gran número de individuos de más 50 años elegidos al azar en esa comunidad, y se quiere tener una idea de la proporción de resultados positivos que se obtendrán. Esto es equivalente a calcular la probabilidad de que la prueba diagnóstica de un resultado positivo en uno de esos individuos.

Es conveniente definir los eventos que usaremos para resolver este problema. Llamemos: R^+ = "el resultado de la prueba es positivo", R^- = "el resultado es negativo", D = "el individuo tiene diabetes" y ND = "el individuo no tiene diabetes".

Conocemos lo siguiente:

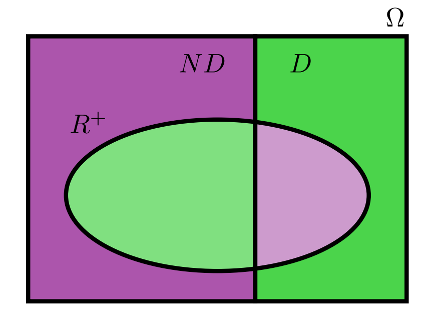
$$\begin{aligned}
\text{Prevalencia } &= P(D) = 0.08, \text{ entonces } P(ND) = 0.92 \\
\text{Sensibildad } &= P(R^+|D) = 0.95, \text{ entonces } P(R^-|D) = 0.05 \\
\text{Especificidad } &= P(R^-|ND) = 0.98, \text{ entonces } P(R^+|ND) = 0.02
\end{aligned}$$

y queremos calcular $P(R^+)$.

En este caso, D y ND son eventos disjuntos y también $D \cup ND = \Omega$, esto significa que constituyen una partición del espacio, que en este caso es toda la población de referencia.

Entonces podemos escribir:

$$\begin{aligned}
R^+ &= R^+ \cap (D \cup ND) \\
&= (R^+ \cap D) \cup (R^+ \cap ND)
\end{aligned}$$



Aplicando la Ley aditiva en (1) y la Regla de la multiplicación en (2):

$$\begin{aligned}
P(R^+) &\overset{(1)}{=} P(R^+ \cap D) + P(R^+ \cap ND) \\
&\overset{(2)}{=} P(R^+|D) \times P(D) + P(R^+|ND) \times P(ND).
\end{aligned}$$

Ahora, reemplazando por los valores, tenemos:

$$P(R^+) = 0.95 \times 0.08 + 0.02 \times 0.92 = 0.0944.$$

■

El procedimiento que utilizamos en este ejemplo es una aplicación del Teorema de la Probabilidad Total.

TEOREMA DE BAYES: Si $A_1, A_2, \ldots, A_n$ representan una partición del espacio muestral Ω, donde $P(A_i) > 0$ para todo i, y sea B un evento cualquiera con $P(B) > 0$, entonces para cualquier $k = 1, \ldots, n$, se cumple:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Demostración: Primero, por la definición de probabilidad condicional tenemos que:

$$P(A_k|B) = \frac{P(A_k \cap B)}{P(B)} \tag{1.3}$$

Segundo, como tenemos las mismas hipótesis que en el Teorema de la Probabilidad Total, podemos

afirmar que:

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i) \tag{1.4}$$

Tercero, si aplicamos la Regla de la multiplicación al numerador de (1.3), tenemos que:

$$P(A_k \cap B) = P(B|A_k)P(A_k) \tag{1.5}$$

Por último, reemplazando en (1.3) los resultados de (1.4) y (1.5), podemos concluir que:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Ejemplo 1.16

Volviendo al Ejemplo 1.15, supongamos que al individuo elegido al azar se le realizó la prueba diagnóstica, y esta dio un resultado positivo, ¿cuál es la probabilidad de que dicho individuo tenga realmente diabetes?

Ahora lo que se desea es calcular $P(D|R^+)$, si aplicamos la definición de probabilidad condicional:

$$P(D|R^+) = \frac{P(D \cap R^+)}{P(R^+)}$$

calculamos $P(D \cap R^+)$ por la Regla de la multiplicación y reemplazamos $P(R^+)$ que ya calculamos previamente, tenemos:

$$P(D|R^+) = \frac{P(R^+|D)P(D)}{P(R^+|D)P(D) + P(R^+|ND)P(ND)}$$

Esto se suele llamar valor predictivo positivo (VPP) de una prueba diagnóstica, es la probabilidad de que el individuo este enfermo dado que la prueba dio un resultado positivo. En nuestro caso:

$$P(D|R^+) = \frac{0.95 \times 0.08}{0.0944} = 0.8051$$

De la misma manera se define el valor predictivo negativo (VPN) de una prueba diagnóstica, que es la probabilidad de que el individuo esté sano dado que el resultado de la prueba fue negativo:

$$P(ND|R^-) = \frac{P(R^-|ND)P(ND)}{P(R^-|D)P(D) + P(R^-|ND)P(ND)}$$

■

Este ejemplo fue una aplicación del Teorema de Bayes.

EJERCICIO 1.8

Resolver los siguientes ejercicios utilizando el Teorema de la Probabilidad Total y el Teorema de Bayes.

1. Una persona toma al azar de una caja uno de los números 1, 2 ó 3, y luego tira un dado

equilibrado tantas veces como indica el número escogido. Después suma el resultado de las tiradas del dado. ¿Cuál es la probabilidad de que obtenga un total de 5?

2. Una compañía utiliza las líneas A_1, A_2 y A_3 para la producción de un microchip. De los microchip fabricados por la línea A_1, el 5 % son defectuosos; de los fabricados por la línea A_2, el 8 % son defectuosos y el 10 % de los fabricados por A_3 son defectuosos. El 50 % de todos los microchips son producidos por A_1, el 30 % por A_2 y el restante por A_3. Se selecciona un microchip al azar.

   a. ¿Cuál es la probabilidad de que el microchip haya sido producido por A_3 y sea defectuoso?

   b. ¿Cuál es la probabilidad de que el microchip sea defectuoso?

   c. Si se observa que es defectuoso, ¿cuál es la probabilidad de que el microchip haya sido producido por A_1?

Referencias

Cramer, H. (1968). *Elementos de la Teoría de Probabilidades y algunas de sus aplicaciones.* Madrid. Ed. Aguilar.

Devore Jay, L. (2001). *Probabilidad y Estadística para Ingeniería y Ciencias.* Ed. Books/Cole Publishing Company.

Feller, W. (1975). *Introducción a la Teoría de Probabilidades y sus Aplicaciones.* Ed. Limusa-Wiley S.A.

Maronna, R. (1995). *Probabilidad y Estadística Elementales para Estudiantes de Ciencias.* Buenos Aires. Ed. Exactas.

Mendenhall, W., Beaver, R. J. & Beaver, B. M. (2006). *Introducción a la Probabilidad y Estadística.* México. Cengage Learning Editores.

Meyer Paul, L. (1970). *Probabilidad y aplicaciones Estadísticas.* Addison-Wesley Iberoamericana.

Parzen, E. (1987). *Teoría Moderna de Probabilidades y sus Aplicaciones.* Ed. Limusa.

Ross, S. M. (1987). *Introduction to Probability and Statistics for Enginees and Scientists.* John Wiley & Sons.

Ross, S. M. (1997). *A first course in Probability.* New Jersey. Pearson Prentice Hall.

Walpole, R. E. & Myers, R. H. (2007). *Probabilidad y Estadística para Ingeniería y Ciencias.* México. Ediciones McGraw-Hill.

CAPÍTULO 2

Variables aleatorias discretas

Variables aleatorias

Al realizar un experimento aleatorio, muchas veces no estamos interesados en el resultado sino en una función del mismo. Por ejemplo, si tiramos dos veces un dado podemos estar interesados en saber cuál es la suma de los resultados de ambas tiradas, cuántas veces salió un valor en particular, cuál es el máximo de los dos valores observados, etc.

En muchos experimentos aleatorios el espacio Ω no es un espacio numérico, entonces nos puede interesar transformar los resultados en valores numéricos.

Podemos lograr ese objetivo definiendo una función que a cada elemento del espacio muestral le haga corresponder un número.

Definición:

Una **variable aleatoria** X es una función que a cada elemento w del espacio muestral Ω, le hace corresponder un número real. Es decir, $X : \Omega \to \mathbb{R}$ si $\omega \in \Omega$, $X(\omega) \in \mathbb{R}$.

Notación

En general abreviaremos *variable aleatoria* escribiendo v.a.

Ejemplo 2.1

Se tira un dado dos veces y se observa $X =$ "el número de veces que sale 1". ■

Ejemplo 2.2

Se tira un dado dos veces y se observa $Y =$ "el máximo de los dos valores". ■

Ejemplo 2.3

Se tira una moneda hasta que sale cara y se define $Z =$ "el número de tiradas necesarias". ■

Ejemplo 2.4

Se administra un nuevo tratamiento a tres personas que padecen una enfermedad, interesa conocer la eficacia de ese tratamiento para lograr la recuperación en una semana (esto también puede considerarse un experimento aleatorio), se observa $V =$ "el número de pacientes, entre los tres tratados, que se recupera en una semana". ■

Ejemplo 2.5

Se elige una persona al azar en una población y se observa $W=$ "peso de la persona elegida". ■

Ejemplo 2.6

Se enciende una lámpara y se observa $T =$ "el tiempo hasta que se quema". ■

Todas las variables definidas en los ejemplos anteriores son variables aleatorias. Ahora, si consideramos el conjunto de valores que puede tomar cada una de ellas vemos que:

$$v_X = \{0, 1, 2\} \qquad v_V = \{0, 1, 2, 3\}$$
$$v_Y = \{1, 2, 3, 4, 5, 6\} \qquad v_W = (0, \infty)$$
$$v_Z = \{1, 2, 3, \ldots\} \qquad v_T = (0, \infty)$$

Los conjuntos de valores v_X, v_Y y v_V son finitos, v_Z es infinito numerable (ya que hay un primer elemento, un segundo elemento, etc.), por otra parte v_T y v_W son infinitos no numerables.

Definición:

Cuando el conjunto de valores (también llamado *rango*) que toma una v.a. es finito o infinito numerable, la variable se denomina **discreta**.

Notación

Sea $a \in \mathbb{R}$ y X una v.a., se utilizará la notación $(X = a)$ para hacer referencia al evento de Ω formado por todos los resultados para los cuales X toma el valor a, y $(X \leq a)$ para el evento formado por todos aquellos resultados para los que X toma valores menores o iguales que a. Esto se puede escribir:

$$(X = a) = \{\omega \in \Omega \text{ tal que } X(\omega) = a\}$$
$$(X \leq a) = \{\omega \in \Omega \text{ tal que } X(\omega) \leq a\}$$

De la misma manera se utilizará la notación: $(X < a)$, $(X > a)$ y $(X \geq a)$.

Ejemplo 2.7

Si se considera el Ejemplo 2.1, el espacio muestral es $\Omega = \{(1,1);(1,2);\ldots;(1,6);(2,1);(2,2);\ldots;(2,6);\ldots;(6,1);(6,2);\ldots;(6,6)\}$ y la variable X definida allí es "el número de veces que sale 1". Podemos definir los eventos:

$$(X = 0) = \{(2,2);(2,3);\ldots;(2,6);(3,2);(3,3);\ldots;(3,6);\ldots;(6,2);(6,3);\ldots;(6,6)\}$$
$$(X = 1) = \{(1,2);(1,3);(1,4);(1,5);(1,6);(2,1);(3,1);(4,1);(5,1);(6,1)\}$$
$$(X = 2) = \{(1,1)\}$$

Si suponemos que el dado es equilibrado y el espacio muestral es equiprobable, con lo cual es fácil ver que:

$$P(X = 0) = 25/36, \qquad P(X = 1) = 10/36 \quad \text{y} \quad P(X = 2) = 1/36$$

■

Ejemplo 2.8

Para la variable Y definida en el Ejemplo 2.2, el espacio muestral Ω es el mismo del

Ejemplo 2.7. Luego podemos definir los eventos:

$$(Y=1)=\{(1,1)\}$$
$$(Y=2)=\{(1,2);(2,1);(2,2)\}$$
$$(Y=3)=\{(1,3);(3,1);(2,3);(3,2);(3,3)\}$$
$$(Y=4)=\{(1,4);(4,1);(2,4);(4,2);(3,4);(4,3);(4,4)\}$$
$$(Y=5)=\{(1,5);(5,1);(2,5);(5,2);(3,5);(5,3);(4,5);(5,4);(5,5)\}$$
$$(Y=6)=\{(1,6);(6,1);(2,6);(6,2);(3,6);(6,3);(4,6);(6,4);(5,6);(6,5);(6,6)\}$$

Si el dado es equilibrado, podemos calcular las probabilidades $P(Y=y)$ para valores de $y=1,2,3,4,5,6$ como:

$$P(Y=1)=1/36 \qquad P(Y=2)=3/36 \qquad P(Y=3)=5/36$$
$$P(Y=4)=7/36 \qquad P(Y=5)=9/36 \qquad P(Y=6)=11/36$$

■

Función de frecuencia de probabilidad

Definición:

Sea X una v.a. discreta y v_X su conjunto de valores. Se define la **función de frecuencia de probabilidad** (o simplemente función de frecuencia) de X como:

$$f(x)=P(X=x) \quad \text{para todos los } x\in v_X$$

La función de frecuencia nos permite calcular probabilidades referidas a la v.a. X:

$$P(X\in A)=\sum_{\substack{x\in A\\ x\in v_X}} f(x), \quad \text{para todo } A\subseteq\mathbb{R} \tag{2.1}$$

En particular si $A=[a,b]$:

$$P(a\le X\le b)=\sum_{\substack{a\le x\le b\\ x\in v_X}} f(x), \quad \text{para todo } a,b\in\mathbb{R} \tag{2.2}$$

PROPIEDAD 2.1: Sea X v.a. discreta, su función de frecuencia f cumple:

- $f(x)\ge 0$, para todo $x\in v_X$
- $\sum_{x\in v_X} f(x)=1$.

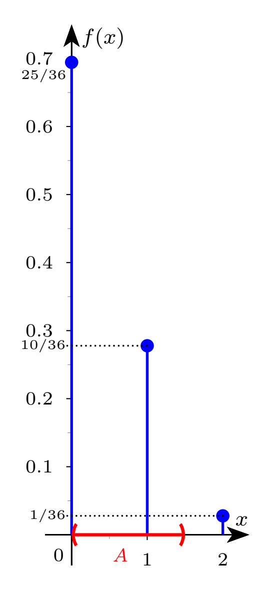
EJERCICIO 2.1

1. Demostrar la Propiedad 2.1.

2. Sea X una v.a. discreta con $v_X = \{-3, 1, 5, 8\}$. ¿Cuáles de las siguientes funciones corresponde a una función de frecuencia de X? Justifique.

   - $f(1) = 0.2,\ f(-3) = 0.3,\ f(5) = 0.5002,\ f(8) = -0.0002.$
   - $f(1) = 1/8,\ f(-3) = 1/2,\ f(5) = 1/8$ y $f(8) = 1/4.$
   - $f(1) = 5/12,\ f(-3) = 1/6,\ f(5) = 1/4$ y $f(8) = 3/4.$

Ejemplo 2.9



La función de frecuencia de la v.a. X definida en el Ejemplo 2.7, está dada por:

x	0	1	2
$f(x)$	25/36	10/36	1/36

Se puede comprobar fácilmente que verifica la Propiedad 2.1. La gráfica de la función de frecuencia se encuentra a la izquierda (por convención las frecuencias se grafican como segmentos verticales).
A partir de esta función se pueden determinar, por (2.1), todas las probabilidades que uno desee. Por ejemplo, si $A = (0, 1.5)$

$$P(X \in A) = P(0 < X < 1.5) = P(X = 1) = f(1) = \frac{10}{36}.$$

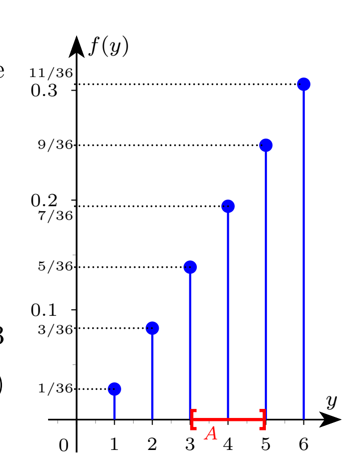
■

Ejemplo 2.10

La función de frecuencia de la v.a. Y hallada en el Ejemplo 2.8 se resume en la siguiente tabla:

y	1	2	3	4	5	6
$f(y)$	1/36	3/36	5/36	7/36	9/36	11/36

La cual también verifica la Propiedad 2.1.
Si se define el evento $A =$ "el máximo de los dos valores está entre 3 y 5 inclusive", este evento puede escribirse como $A = (3 \leq Y \leq 5)$ y su probabilidad se calcula como:



$$P(A) = P(3 \le Y \le 5) = \sum_{3\le y\le 5} f(y) = \sum_{y=3}^{5} f(y) \qquad \big(\text{por } (2.2)\big)$$
$$= f(3) + f(4) + f(5) = 5/36 + 7/36 + 9/36 = 21/36.$$

■

Función de distribución o función de distribución acumulada

Definición:

La **función de distribución** o **función de distribución acumulada** de una v.a. X se define como:

$$F(x) = P(X \le x) \quad \text{para todo } x \in \mathbb{R}. \tag{2.3}$$

Se puede comprobar fácilmente que la función de distribución cumple:

- es una función no decreciente: si $a, b \in \mathbb{R}$ y $a < b$ entonces $F(a) \le F(b)$
- toma valores entre 0 y 1

Notación

En general abreviaremos *función de distribución acumulada* escribiendo fda.

PROPIEDAD 2.2: Sea F la fda de la v.a. X, sean $a, b \in \mathbb{R}$ tales que $a < b$, entonces se cumple:

$$P(a < X \le b) = F(b) - F(a)$$

Demostración: Como $a < b$ entonces podemos escribir $(X \le b) = (X \le a) \cup (a < X \le b)$ y estos dos eventos son disjuntos entonces por **(A3)**, la Ley aditiva:

$$P(X \le b) = P(X \le a) + P(a < X \le b),$$

por lo tanto, despejando y aplicando (2.3), obtenemos que $P(a < X \le b) = F(b) - F(a)$.

Aclaración

Notar que en la definición anterior y en la Propiedad 2.2 no estamos aclarando que la v.a. X sea una v.a. discreta. Es decir, la definición y esta propiedad son válidas tanto para v.a. discretas como para continuas.

Cuando la v.a. X es discreta, la fda se calcula como:

$$F(x) = P(X \leq x) = \sum_{\substack{k \leq x \\ k \in v_X}} f(k) \tag{2.4}$$

Entonces, la función de distribución de una v.a. discreta es escalonada, con saltos en los valores que toma la variable y constante en el resto. Notar que la magnitud del salto es igual a la función de frecuencia en este valor (ver el siguiente ejemplo).

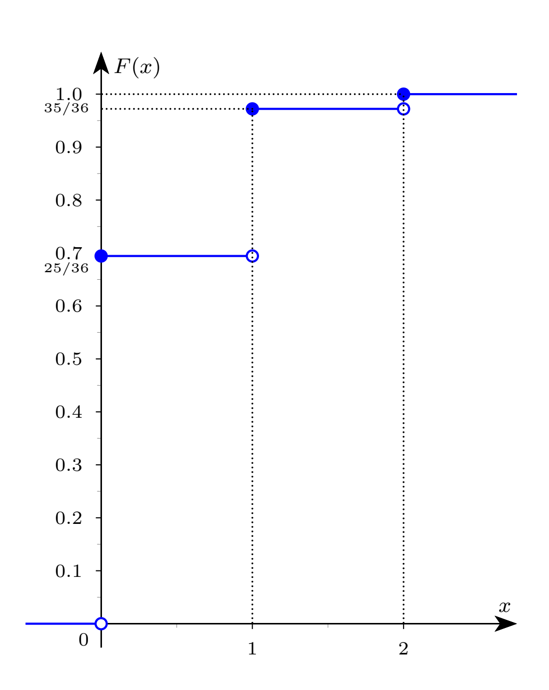
Ejemplo 2.11

Se puede calcular la fda de la v.a. X del Ejemplo 2.9, utilizando (2.4), de la siguiente manera:

- Si $x < 0$, $F(x) = P(X \leq x) = \sum_{k \leq x} f(k) = 0$
- Si $0 \leq x < 1$, $F(x) = P(X \leq x) = \sum_{k \leq x} f(k) = f(0) = \frac{25}{36}$
- Si $1 \leq x < 2$, $F(x) = P(X \leq x) = \sum_{k \leq x} f(k) = f(0) + f(1) = \frac{35}{36}$
- Si $x \geq 2$, $F(x) = P(X \leq x) = \sum_{k \leq x} f(k) = f(0) + f(1) + f(2) = 1$

Resumiendo, la función de distribución para la v.a X es:

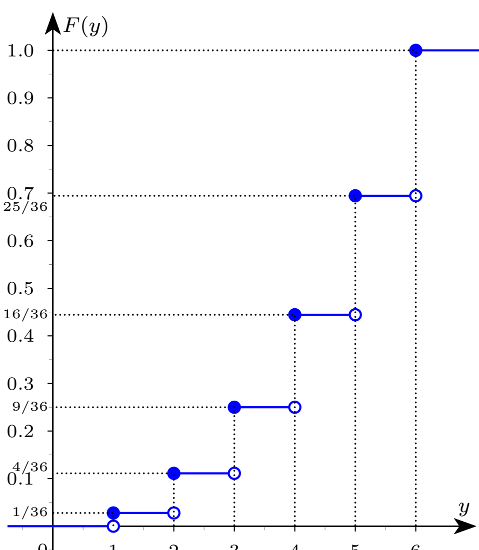
$$F(x) = \begin{cases} 0 & si \quad x < 0 \\ 25/36 & si \quad 0 \leq x < 1 \\ 35/36 & si \quad 1 \leq x < 2 \\ 1 & si \quad x \geq 2 \end{cases}$$



Ejemplo 2.12

De la misma manera se puede calcular la fda de la v.a. Y del Ejemplo 2.10. Resumiendo, la función de distribución para la v.a Y es:

$$F(y) = \begin{cases} 0 & si \quad y < 1 \\ 1/36 & si \quad 1 \leq y < 2 \\ 4/36 & si \quad 2 \leq y < 3 \\ 9/36 & si \quad 3 \leq y < 4 \\ 16/36 & si \quad 4 \leq y < 5 \\ 25/36 & si \quad 5 \leq y < 6 \\ 1 & si \quad y \geq 6 \end{cases}$$



Luego, la probabilidad de cualquier evento que se relacione con el máximo de las dos tiradas puede calcularse usando esta función de distribución. Por ejemplo, sean los eventos:

$A =$ “el máximo de las dos tiradas es a lo sumo 3”,

$B =$ “el máximo de las dos tiradas es 4” y

$C =$ “el máximo es mayor que 2 y menor que 5”.

Entonces sus probabilidades son:

$$P(A) = P(Y \leq 3) = F(3) = 9/36$$

$$P(B) = P(Y = 4) = P(Y \leq 4) - P(Y \leq 3) = F(4) - F(3) = 7/36$$

$$P(C) = P(2 < Y < 5) = P(2 < Y \leq 4) = F(4) - F(2) = 1/3$$

■

EJERCICIO 2.2

Calcular para cada una de las siguientes v.a. X la función de frecuencia y la fda. Graficar ambas funciones.

1. Supongamos un juego donde se tira un dado y usted gana \$12 si en el dado sale 6 y pierde \$3 si sale otro número. Sea la v.a. $X =$ “ganancia en este juego” y $v_X = \{-3, 12\}$, en donde -3 refleja que se han perdido \$3, lo que representa una ganancia negativa.

2. Cinco pelotas numeradas del 1 al 5 se colocan en una urna. Se seleccionan dos de ellas al azar. Sea la v.a. $X =$ “el mayor número obtenido”.

3. Una pieza de equipo electrónico contiene 6 chips de computadora, dos de los cuales son defectuosos. Al azar se seleccionan tres chips, se retiran del equipo y se inspeccionan. Sea la v.a. $X =$ “el número de chips defectuosos observados”.

Variables aleatorias independientes

Tal como definimos el concepto de independencia entre dos eventos A y B, definimos la independencia de v.a. Lo que queremos decir intuitivamente es que si X e Y son v.a. independientes, el resultado de una de ellas no influye en el resultado de la otra.

Definición:

Las v.a. X, Y son **independientes** si y sólo si para todo a, $b \in \mathbb{R}$, los eventos $(X \leq a)$ e $(Y \leq b)$ son independientes.



Observación:

En particular, para v.a. discretas, se puede decir que X e Y son independientes si y sólo si para todo a, $b \in \mathbb{R}$, los eventos $(X = a)$ e $(Y = b)$ son independientes.

Esta noción será útil para representar los resultados de experimentos que no se influyen mutuamente.

Ejemplo 2.13

Se arrojan dos dados equilibrados a la vez, uno de color rojo y el otro verde. Considerar las v.a.:

$$X = \text{"el número del dado rojo"}.$$

$$Y = \text{"el número del dado verde"}.$$

$$Z = \text{"la suma de los dos dados"}.$$

¿Las v.a. X e Y son independientes? ¿Las v.a. X y Z son independientes?

Primero veamos las funciones de frecuencia de estas tres variables. El espacio muestral en este caso es: $\Omega = \{(1,1);(1,2);\ldots;(1,6);(2,1);(2,2);\ldots;(2,6);\ldots;(6,1);(6,2);\ldots;(6,6)\}$, donde la primer coordenada es el resultado del dado rojo y la segunda es el resultado del dado verde. Luego podemos definir los eventos:

$$(X = a) = \{(a,1);(a,2);(a,3);(a,4);(a,5);(a,6)\}, \text{ para todo } a \in \{1,2,\ldots,6\}$$

$$(Y = b) = \{(1,b);(2,b);(3,b);(4,b);(5,b);(6,b)\}, \text{ para todo } b \in \{1,2,\ldots,6\}$$

$$(Z = 2) = \{(1,1)\}$$

$(Z = 3) = \{(1,2);(2,1)\}$

$(Z = 4) = \{(1,3);(2,2);(3,1)\}$

$\vdots$

$(Z = 10) = \{(4,6);(5,5);(6,4)\}$

$(Z = 11) = \{(5,6);(6,5)\}$

$(Z = 12) = \{(6,6)\}$

Es claro que $f_X(a) = P(X = a) = 6/36 = 1/6$ para todo $a \in \{1,2,\ldots,6\}$, $f_Y(b) = P(Y = b) = 6/36 = 1/6$ para todo $b \in \{1,2,\ldots,6\}$ y

z	2	3	4	5	6	7	8	9	10	11	12
$f_Z(z)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Para todo $a, b \in \{1,2,\ldots,6\}$, $P\big((X = a) \cap (Y = b)\big) = 1/36$ y $P(X = a) \times P(Y = b) = 1/6 \times 1/6 = 1/36$, es decir, $(X = a)$ e $(Y = b)$ son independientes. Por lo tanto, las v.a. X e Y son v.a. independientes.

Por otro lado, $P\big((X = 1) \cap (Z = 2)\big) = P\big(\{(1,1)\}\big) = 1/36$ y $P(X = 1) \times P(Z = 2) = 1/6 \times 1/36 = 1/216 \neq 1/36$, es decir, $(X = 1)$ y $(Z = 2)$ no son independientes. Por lo tanto, las v.a. X y Z no son v.a. independientes.

■

Valor esperado o media

El valor esperado de una variable aleatoria (llamado también *esperanza matemática*, *valor medio*, o *media*) es el promedio pesado de los valores que toma, en donde cada valor recibe un peso igual a su probabilidad. La media es una medida de centralidad, es decir, nos da un centro alrededor del cual se distribuyen los valores de la v.a.

Definición:

Sea X una v.a. discreta con valores en el conjunto v_X y función de probabilidad f, se define el **valor esperado** de X como:

$$E(X) = \sum_{x \in v_X} x f(x), \tag{2.5}$$

si se cumple que $\sum_{x \in v_X} |x| f(x) < \infty$. Si esta suma diverge se dice $E(X)$ no existe.

El significado intuitivo del valor esperado es el siguiente: imaginemos que el experimento se repite un gran número N de veces, y se toma el promedio de los valores de X observados en cada

repetición, entonces $E(X)$ es el límite de esos promedios cuando N tiende a infinito.

Ejemplo 2.14

Volvamos al juego del Ejercicio 2.2 en el que se tira un dado y usted gana \$12 si en el dado sale 6 y pierde \$3 si sale otro número. ¿Jugaría usted a este juego? ¿Esperaría ganar?

Veamos: primero definamos la v.a. X = "ganancia en este juego", donde $v_X = \{-3, 12\}$.

Ahora, la función de probabilidad de esta v.a., que se obtuvo en el ejercicio, es:

$\boldsymbol{x}$	-3	12
$\boldsymbol{f(x)}$	5/6	1/6

Como en este caso v_X es finito, sabemos que existe la esperanza y la podemos calcular como:

$$E(X) = -3 \times 5/6 + 12 \times 1/6 = -3/6 = -0.5$$

Luego, el valor que uno espera ganar es -0.5. Esto significa que si usted jugara muchas veces a este juego, algunas veces ganaría, otras perdería, pero el promedio final es negativo, a la larga no espere ganar!!!

■

Valor esperado o media de una función de una v.a.

Si queremos calcular por definición la esperanza de una v.a. Y, que es función de una v.a. discreta X, deberíamos calcular su función de frecuencia. Pero si se conoce de antemano la función de frecuencia de X, la siguiente proposición nos permite calcular la media de Y de una manera más sencilla.

PROPOSICIÓN 2.1: Sea X una v.a. discreta con valores en el conjunto v_X y función de frecuencia f y $h : \mathbb{R} \to \mathbb{R}$ una función cualquiera, entonces $Y = h(X)$ es una v.a. cuya media se calcula como:

$$E(Y) = E\big(h(X)\big) = \sum_{x \in v_X} h(x) f(x) \tag{2.6}$$

si se cumple que $\sum_{x \in v_X} |h(x)| f(x) < \infty$. Si esta suma diverge se dice $E(Y)$ no existe.

Aceptamos este resultado sin demostración.

Una consecuencia inmediata de la proposición anterior, es que el valor medio tiene la siguiente propiedad:

PROPIEDAD DE LINEALIDAD DE LA ESPERANZA: Sea X una v.a. con media $E(X)$ y sean a y b números reales, entonces

$$E(aX + b) = aE(X) + b. \tag{2.7}$$

Demostración: Si X es discreta con valores en el conjunto v_X y función de probabilidad f, la demostración de esta propiedad es simple, utilizando la Proposición anterior con $h(X) = aX + b$, tenemos que:

$$\begin{aligned} E(aX + b) &= \sum_{x \in v_X} (ax + b) f(x) \\ &= \sum_{x \in v_X} \big(axf(x) + bf(x)\big) && \big(\text{distributiva en el sumando}\big) \\ &= a \sum_{x \in v_X} xf(x) + b \sum_{x \in v_X} f(x) && \big(\text{distributiva y factor común en la sumatoria}\big) \\ &= aE(X) + b && \big(\text{por (2.5) y por la Propiedad 2.1}\big). \end{aligned}$$

EJERCICIO 2.3

Sea X una v.a. discreta con función de frecuencia de X:

x	0	1	2	3	4
$f(x)$	0.08	0.15	0.45	0.27	0.05

1. Calcular la $E(\sqrt{X})$.
2. Calcular la $E(-2\sqrt{X} + 3.5)$.

Varianza y desviación típica

Ya definimos que la media es una medida de centralidad. Ahora, vamos a definir un parámetro que nos da una idea de la dispersión de los valores de X alrededor de su valor medio.

Definición:

Sea X una v.a. que tiene media $E(X)$, se define la **varianza** de X como:

$$var(X) = E\Big[\big(X - E(X)\big)^2\Big] \tag{2.8}$$

cuando dicha esperanza existe. Y se define la **desviación típica** (o *estándar*) como:

$$dt(X) = \sqrt{var(X)} \tag{2.9}$$

La $var(X)$ (o $V(X)$) se expresa en las unidades de X al cuadrado, pero $dt(X)$ se expresa en las mismas unidades que X.

La siguiente propiedad nos permite una forma práctica de calcular la varianza.

PROPIEDAD 2.3: La definición de $var(X)$ es equivalente a:

$$var(X) = E(X^2) - \big(E(X)\big)^2$$

Demostración: Sea X una v.a. discreta con valores en el conjunto v_X, función de probabilidad f y llamamos $E(X) = \mu$:

$$\begin{aligned}
var(X) &= E\big[(X-\mu)^2\big] = E(X^2 - 2X\mu + \mu^2) && \text{(desarrollo del cuadrado)}\\
&= \sum_{x\in v_X} (x^2 - 2x\mu + \mu^2) f(x) && \big(\text{por (2.6)}\big)\\
&= \sum_{x\in v_X} x^2 f(x) - 2\mu \sum_{x\in v_X} x f(x) + \mu^2 \sum_{x\in v_X} f(x) && \text{(distributiva y factor común)}\\
&= E(X^2) - 2\mu^2 + \mu^2 && \big(\text{por (2.5), (2.6) y Propiedad 2.1}\big)\\
&= E(X^2) - \mu^2
\end{aligned}$$

PROPIEDAD 2.4: Sea X una v.a. y sean a y b números reales. Entonces:

$$var(aX+b) = a^2 var(X) \tag{2.10}$$
$$dt(aX+b) = |a|\, dt(X)$$

Demostración: Por la definición dada en (2.8), tenemos que:

$$var(aX+b) = E\Big[\big((aX+b) - E(aX+b)\big)^2\Big]$$

Luego,

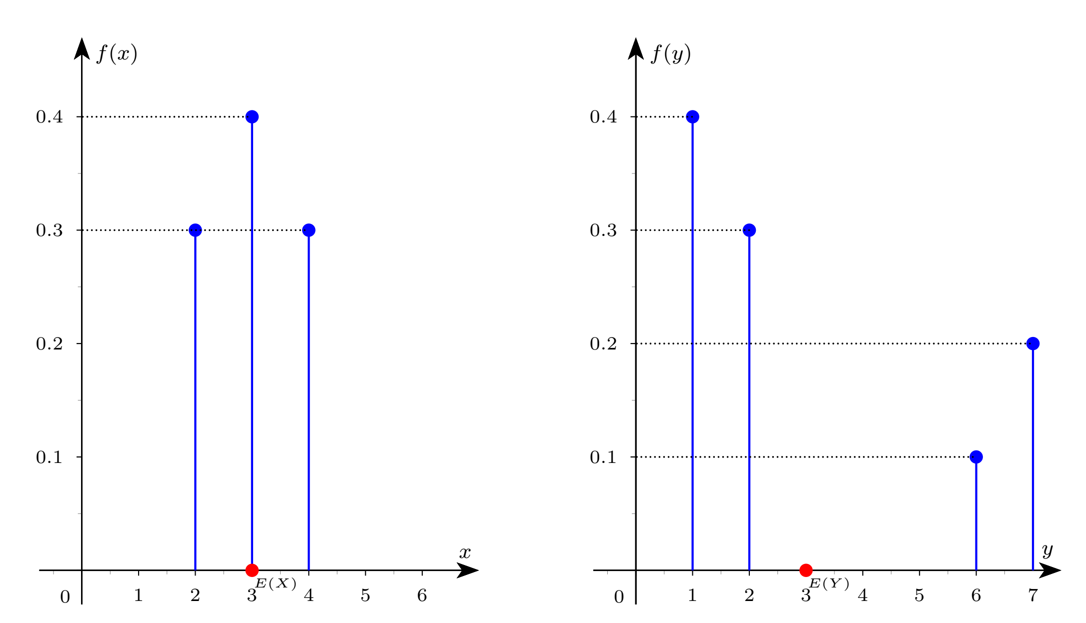
$$\begin{aligned}
E\Big[\big((aX+b) - E(aX+b)\big)^2\Big] &= E\Big[(aX + \not{b} - aE(X) - \not{b})^2\Big] && \big(\text{por (2.7)}\big)\\
&= E\Big[\big(a(X - E(X))\big)^2\Big] && \big(\text{factor común}\big)\\
&= E\Big[a^2\big(X - E(X)\big)^2\Big]\\
&= a^2 E\Big[\big(X - E(X)\big)^2\Big] && (*)\\
&= a^2 var(X) && \big(\text{por (2.8)}\big)
\end{aligned}$$

En $(*)$ observar que si $Y = \big(X - E(X)\big)^2$ es una v.a. y por (2.7) tenemos que $E(a^2Y) = a^2E(Y)$.

Por último:

$$\begin{aligned} dt(aX+b) &= \sqrt{var(aX+b)} && \big(\text{por } (2.9)\big) \\ &= \sqrt{a^2 var(X)} && \big(\text{por } (2.10)\big) \\ &= \sqrt{a^2}\sqrt{var(X)} && \big(\text{distributiva de la raíz}\big) \\ &= |a|\, dt(X) && \big(\text{por } (2.9)\big) \end{aligned}$$

Ejemplo 2.15



Aún cuando ambas distribuciones ilustradas tienen la misma media ($E(X) = E(Y) = 3$), la distribución de la v.a. Y tiene mayor dispersión o variabilidad que la v.a. X. Calcule las varianzas en ambos casos y compare.

■

Algunas variables aleatorias discretas

Se pueden hallar ejemplos de v.a. discretas en numerosas aplicaciones cotidianas y en casi todas las disciplinas. No obstante, hay dos distribuciones de probabilidad discretas que sirven para modelizar un gran número de estas aplicaciones, la distribución de probabilidad binomial y la distribución de Poisson, las cuales estudiaremos en esta sección.

Distribución binomial

Definición:

Un **experimento binomial** es el que cumple las siguientes condiciones:

1. El experimento consiste en n repeticiones idénticas de un ensayo que toma dos resultados posibles, que se denotan éxito (E) y fracaso (F).
2. Las repeticiones son independientes, lo que significa que el resultado de cualquier repetición particular no influye en el resultado de ninguna otra.
3. La probabilidad de éxito es constante en cada repetición del ensayo, esta probabilidad se denota con $P(E) = p$. Se deduce así, que la probabilidad de fracaso será igual a $P(F) = 1 - p$.

Con este tipo de experimentos se asocia la v.a. binomial.

Ejemplo 2.16

Supongamos que en un hospital hay 3 pacientes internados con determinada enfermedad, a los cuales se les aplica el mismo tratamiento (estos individuos no son parientes). Supongamos que la probabilidad de que un individuo se recupere en una semana de tratamiento es 0.8 ($p = 0.8$). Sea Y la variable aleatoria que cuenta el número de individuos que se recuperan en una semana de tratamiento entre los 3. Los posibles resultados y sus respectivas probabilidades se resumen en la siguiente tabla, donde S y N indican que el individuo se recupera y no se recupera:

Ω: resultados posibles de la evolución de 3 pacientes	**Probabilidad del resultado obtenido (Se utiliza independencia)**	**Valores de Y**
(S, S, S)	$0.8 \times 0.8 \times 0.8 = 0.8^3 = 0.8^3 \times (1-0.8)^0$	3
(S, N, N)	$0.8 \times (1-0.8) \times (1-0.8) = 0.8^1 \times (1-0.8)^2$	1
(N, S, N)	$(1-0.8) \times 0.8 \times (1-0.8) = 0.8^1 \times (1-0.8)^2$	1
(N, N, S)	$(1-0.8) \times (1-0.8) \times 0.8 = 0.8^1 \times (1-0.8)^2$	1
(S, S, N)	$0.8 \times 0.8 \times (1-0.8) = 0.8^2 \times (1-0.8)^1$	2
(S, N, S)	$0.8 \times (1-0.8) \times 0.8 = 0.8^2 \times (1-0.8)^1$	2
(N, S, S)	$(1-0.8) \times 0.8 \times 0.8 = 0.8^2 \times (1-0.8)^1$	2
(N, N, N)	$(1-0.8) \times (1-0.8) \times (1-0.8) = 0.8^0 \times (1-0.8)^3$	0

Si nos interesa únicamente saber cuántos pacientes se recuperan en la primera semana de tratamiento (el valor de Y), y las respectivas probabilidades, se puede resumir aún más:

Valor de Y	**Probabilidad**
0	$1 \times 0.8^0 \times (1-0.8)^3$
1	$3 \times 0.8^1 \times (1-0.8)^2$
2	$3 \times 0.8^2 \times (1-0.8)^1$
3	$1 \times 0.8^3 \times (1-0.8)^0$

Veamos con detalle cómo se llegó a estos resultados tomando uno de los casos como ejemplo:

$$\begin{aligned} f(1) = P(Y=1) &= P\{(S,N,N),(N,S,N),(N,N,S)\} \\ &= P\{(S,N,N)\} + P\{(N,S,N)\} + P\{(N,N,S)\} \qquad \text{(por ser disjuntos)} \\ &= 0.8^1 \times (1-0.8)^2 + 0.8^1 \times (1-0.8)^2 + 0.8^1 \times (1-0.8)^2 = 3 \times 0.8^1 \times (1-0.8)^2 \end{aligned}$$

Generalizando, la función de frecuencia de la v.a. Y es la que se muestra en el siguiente cuadro:

Valor de Y	**Probabilidad**
0	$1 \times 0.8^0 \times (1-0.8)^3 = \binom{3}{0} \times 0.8^0 \times (1-0.8)^{3-0}$
1	$3 \times 0.8^1 \times (1-0.8)^2 = \binom{3}{1} \times 0.8^1 \times (1-0.8)^{3-1}$
2	$3 \times 0.8^2 \times (1-0.8)^1 = \binom{3}{2} \times 0.8^2 \times (1-0.8)^{3-2}$
3	$1 \times 0.8^3 \times (1-0.8)^0 = \binom{3}{3} \times 0.8^3 \times (1-0.8)^{3-3}$

■

Definición:

El número total de éxitos observados entre los n ensayos de un experimento binomial, es una **variable aleatoria binomial** con parámetros n y p.

Los valores que puede tomar esta variable son: $v_X = \{0, 1, 2, \ldots, n\}$.

La función de frecuencia de X es:

$$f(k) = P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in v_X \tag{2.11}$$

donde $\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$.

Notación

Si X es una v.a. binomial con parámetros n y p, lo denotaremos como $X \sim B(n,p)$.

Para demostrar que la expresión (2.11) representa una función de frecuencia legítima se debe verificar la Propiedad 2.1, es decir:

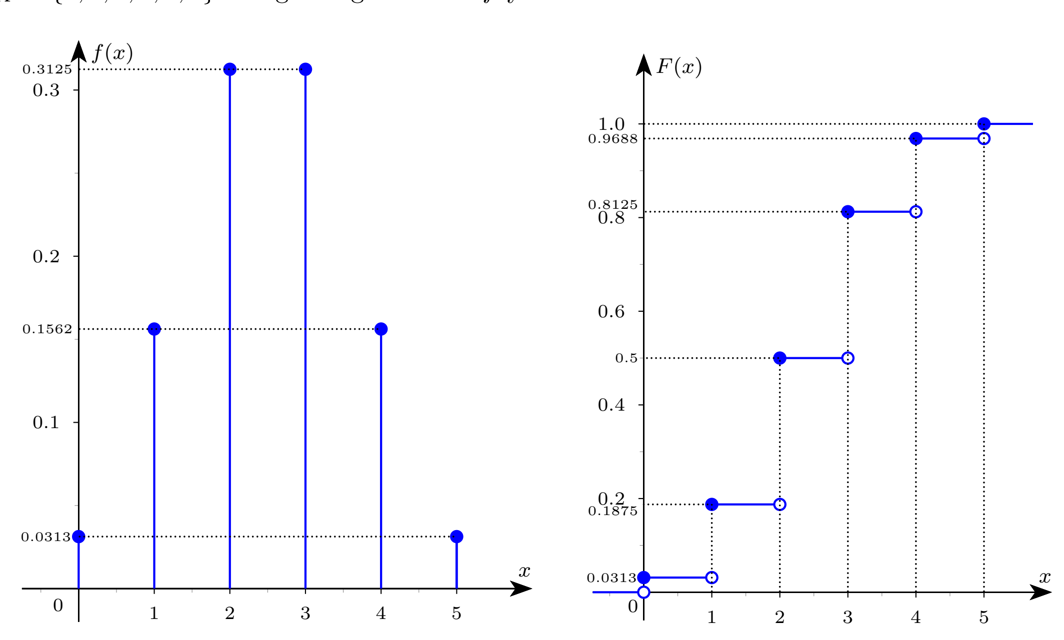
- $f(k) = \binom{n}{k} p^k (1-p)^{n-k} \geq 0$, para todo $k \in v_X$

- $\sum_{k=0}^{n} f(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$ (en este caso se usa el Teorema del binomio de Newton).

Ejemplo 2.17

Se arroja cinco veces una moneda equilibrada. Se desea calcular la función de frecuencia y de distribución del número de caras en las cinco tiradas.

Definimos la v.a. $X=$ "número de caras en las 5 tiradas". Como $X \sim B(5, 0.5)$ entonces $v_X = \{0, 1, 2, 3, 4, 5\}$. Luego las gráficas de f y F son:



PROPOSICIÓN 2.2: Si $X \sim B(n, p)$, entonces:

- $E(X) = np$
- $V(X) = np(1-p)$
- $dt(X) = \sqrt{np(1-p)}$

Esta proposición se puede demostrar utilizando la definición de esperanza y varianza para una v.a. discreta, y recordando además, el Teorema del binomio de Newton.

Ejemplo 2.18

Para la v.a. X del Ejemplo 2.17, se tiene que $E(X) = 5 \times 0.5 = 2.5$, $V(X) = 5 \times 0.5 \times (1 - 0.5) = 1.25$ y $dt(X) = \sqrt{1.25} = 1.1180$.

Para la v.a. $Y \sim B(3, 0.8)$, del Ejemplo 2.16, tenemos que $E(Y) = 3 \times 0.8 = 2.4$, $V(Y) = 3 \times 0.8 \times (1 - 0.8) = 0.48$ y $dt(Y) = \sqrt{0.48} = 0.6928$.

■

EJERCICIO 2.4

De una urna que contiene una bola blanca y nueve bolas negras, se hacen cinco extracciones sucesivas con reemplazo. Llamamos X al número de bolas blancas obtenidas en las cinco extracciones.

1. Indicar la distribución de probabilidad de X, su rango y cuales son sus parámetros.
2. ¿Cuál es la probabilidad de que se saquen exactamente dos bolas blancas?
3. ¿Cuál es la probabilidad de obtener al menos dos bolas blancas?
4. ¿Cuál es la probabilidad de que se saque no más de tres bolas blancas?

Definición:

Si $n = 1$, o sea $X \sim B(1, p)$, la v.a. X toma sólo los valores: 1 con probabilidad p y 0 con probabilidad $1 - p$. En este caso se dice que X tiene distribución de **Bernoulli**.

Resumimos en una tabla la función de frecuencia de X:

$\boldsymbol{x}$	0	1
$\boldsymbol{f(x)}$	$1-p$	p

Observación:



En general, cuando se realizan extracciones con reposición, como en el Ejercicio 2.4, estamos en presencia de un experimento binomial.

Por el contrario, si las extracciones son sin reposición, el resultado de cada extracción depende de las anteriores, de modo que no vale la hipótesis de independencia y por ende no es un experimento binomial. Sin embargo, si la población es grande y la muestra extraída no supera el 5 % del tamaño de la población, cada extracción puede considerarse "prácticamente" independiente de las anteriores y es posible analizar el experimento como binomial. En consecuencia, la v.a. número de éxitos en esas extracciones puede pensarse como binomial.

EJERCICIO 2.5

En cada caso, indicar si el experimento puede ser considerado binomial recordando las condiciones que deberían cumplirse y justificando correctamente.

1. Se tiene una urna con 15 bolillas blancas y 5 verdes. Se extraen al azar y con reemplazo 3 bolillas y se observa si son blancas.

2. Se tiene una urna con 15 bolillas blancas y 5 verdes. Se extraen al azar y sin reemplazo 3 bolillas y se observa si son blancas.

3. Se realizan tres extracciones sin reemplazo de una urna que contiene 1500 bolillas blancas y 500 verdes, interesa observar si se seleccionaron bolillas blancas.

Distribución de Poisson

Definición:

Se dice que una v.a. X tiene **distribución de Poisson** con parámetro λ ($\lambda > 0$) si su función de frecuencia es:

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad \text{para } x = 0, 1, 2, \ldots \tag{2.12}$$

Notación

Si la v.a. X tiene distribución de Poisson con parámetro λ, lo denotaremos como: $X \sim P(\lambda)$.

La distribución de Poisson sirve para modelizar el número X de eventos que ocurren aleatoriamente en el tiempo o en una región. A continuación veamos algunos ejemplos de experimentos en los cuales la variable aleatoria puede ser modelizada con distribución de Poisson:

- El número de llamadas recibidas por un conmutador durante un tiempo determinado.
- El número de bacterias por volumen de fluido.
- El número de llegadas de clientes al mostrador de una caja de pago en un tiempo determinado.
- El número de descomposturas de una máquina durante cierto día.
- El número de accidentes de tránsito en un cruce dado durante un tiempo establecido.
- El número de árboles de determinada especie distribuidos aleatoriamente en un área.

Algunos de estos ejemplos son procesos temporales, interesa conocer cuántas veces ocurre un evento en un intervalo de tiempo, y otros son procesos espaciales, interesa conocer cuántos “puntos” hay en un volumen o un área.

Definición:

Se denomina **proceso temporal de Poisson** cuando cumple con las siguientes características:

- **Invariancia:** las condiciones no cambian en el tiempo.
- **Falta de memoria:** lo que sucede en el intervalo de tiempo $[0, t)$ no influye en lo que suceda en el intervalo $[s, r)$ para $r > s > t$.
- **Sucesos aislados:** la probabilidad de que en un intervalo de tiempo muy corto ocurra más de una vez el evento, es despreciable comparada con la probabilidad de que ocurra una vez o ninguna.

Para un proceso de este tipo, si X_t es la v.a. que mide el número de veces que ocurre el evento en un intervalo de tiempo de longitud t, puede verse que X_t es una variable aleatoria discreta cuya función de frecuencia está dada por:

$$f(x) = e^{-c\times t}\frac{(c\times t)^x}{x!} \quad \text{para } x = 0, 1, 2, \ldots$$

Comparando con la expresión (2.12), se puede ver que X_t tiene distribución de Poisson con parámetro $\lambda_t = c \times t$, donde c es una constante positiva que indica la cantidad de veces que ocurre el evento de interés por unidad de tiempo, c se llama **tasa de ocurrencia del proceso**.

Ejemplo 2.19

Llegan clientes a un mostrador de un negocio con una distribución de Poisson a una tasa de 5 por hora. Si queremos saber cuál es la probabilidad de que no lleguen más de tres clientes en una hora, definimos la v.a. X_1 ="cantidad de clientes que llegan al mostrador en una hora". Entonces $X_1 \sim P(\lambda_1)$, pues $\lambda_1 = 5 \times 1$. Así, la probabilidad pedida es:

$$P(X_1 \leq 3) = F(3) = 0.2650 \qquad \text{(por Tabla)}$$

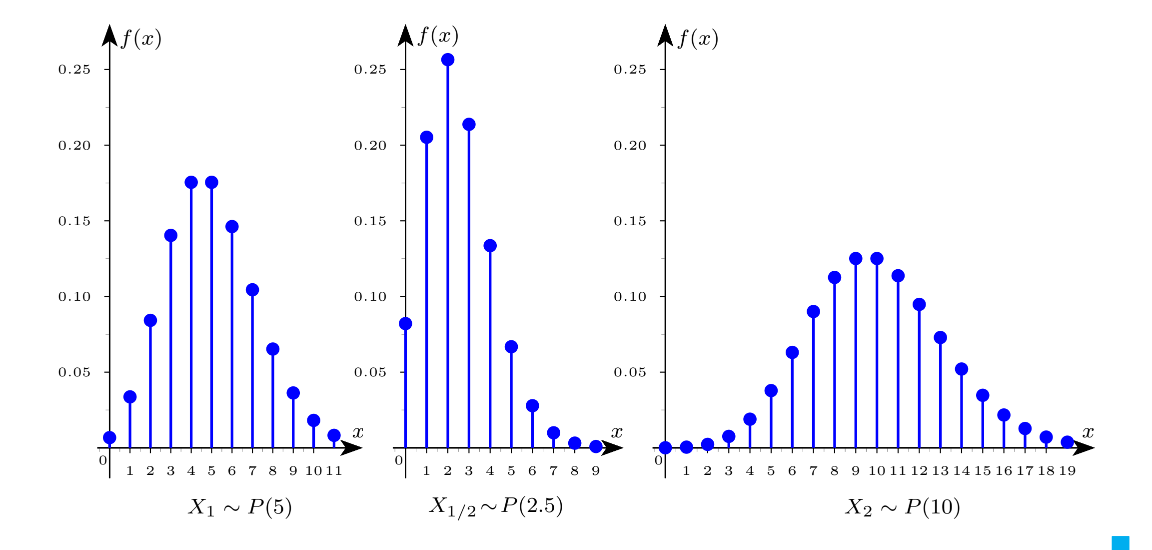
Sin embargo, si queremos calcular la probabilidad de que lleguen al menos 6 clientes en dos horas, no podemos utilizar la v.a. X_1 antes definida, tendremos que redefinirla, ya que el intervalo de tiempo ahora es de 2 hs. Luego, X_2 ="cantidad de clientes que llegan al mostrador en dos horas", $X_2 \sim P(\lambda_2)$, ya que $\lambda_2 = 5 \times 2 = 10$. El cálculo de la probabilidad pedida es:

$$P(X_2 \geq 6) = 1 - P(X_2 < 6) = 1 - P(X_2 \leq 5) = 1 - F(5) = 1 - 0.0671 = 0.9329 \quad \text{(por Tabla)}$$

Por último, si queremos calcular la probabilidad de que lleguen exactamente 5 clientes en media hora, $X_{1/2}$ ="cantidad de clientes que llegan al mostrador en media hora", $X_{1/2} \sim P(2.5)$ y

$$P(X_{1/2} = 5) = e^{-2.5}\frac{2.5^5}{5!} = 0.0668$$

Las gráficas de la función de frecuencia para las v.a. X_1, $X_{1/2}$ y X_2 son, respectivamente:



Definición:

Se denomina **proceso espacial de Poisson** cuando cumple con las siguientes características:

- **Homogeneidad espacial:** la probabilidad de que un punto este en una región dada, sólo depende del tamaño de esa región (área o volumen) y no de su forma o posición.
- **No interacción:** lo que ocurre en una región es independiente de lo que ocurre en otra, si no se superponen.

La v.a. X_a que mide el número de "puntos" en una región de área o volumen a, tiene distribución de Poisson con parámetro $\lambda_a = c \times a$, donde c se interpreta como la tasa de ocurrencia del proceso.

Ejemplo 2.20

La distribución de plantas de cierta especie en una zona sigue un proceso de Poisson con una tasa de 5 plantas por metro cuadrado. Si deseamos calcular la probabilidad de no hallar plantas en un área cuadrada de 1 metro de lado, definimos la v.a. X_1 = "número de plantas en una región cuadrada de área 1 m^2", donde $X_1 \sim P(\lambda_1)$ con $\lambda_1 = 5 \times 1$. Es decir, $X_1 \sim P(5)$ y la probabilidad pedida es $P(X_1 = 0) = e^{-5} \times \frac{5^0}{0!} = 0.0067$.

Ahora, ¿de qué medida debe ser tomado el radio r de una región circular de muestreo para que la probabilidad de hallar al menos una planta de esa especie sea por lo menos 0.99? Necesitamos definir otra v.a. X_a = "número de plantas en una región circular de área $a\,m^2$", donde $X_a \sim P(\lambda_a)$ y $\lambda_a = c \times a$, entonces el planteo es

$$P(X_a > 0) \geq 0.99 \tag{2.13}$$

Si la región de muestreo es circular de radio r, el área de esa región es $a = \pi \times r^2$, y la v.a.

X_a que mide el número de plantas en esa región tendrá distribución de Poisson con parámetro $\lambda_a = c \times a = 5 \times (\pi \times r^2)$, entonces,

$$
\begin{aligned}
P(X_a > 0) = 1 - P(X_a \leq 0) = 1 - P(X_a = 0) \\
= 1 - e^{-5\times\pi\times r^2} \frac{(5 \times \pi \times r^2)^0}{0!} = 1 - e^{-5\times\pi\times r^2}.
\end{aligned}
$$

Luego, si reemplazamos en (2.13), obtenemos:

$$
\begin{aligned}
1 - e^{-5\times\pi\times r^2} &\geq 0.99 \\
0.01 - e^{-5\times\pi\times r^2} &\geq 0 && \text{(restando de ambos lados 0.99)} \\
0.01 &\geq e^{-5\times\pi\times r^2} && \text{(sumando de ambos lados } e^{-5\times\pi\times r^2}\text{)} \\
\ln(0.01) &\geq \ln(e^{-5\times\pi\times r^2}) && \text{(aplicando de ambos lados la función } \ln\text{)} \\
\ln(0.01) &\geq -5 \times \pi \times r^2 && \text{(por propiedad de función inversa)} \\
\frac{\ln(0.01)}{-5 \times \pi} &\leq r^2 && \text{(dividiendo en ambos lados por } -5 \times \pi\text{)} \\
\left[\frac{-\ln(0.01)}{5 \times \pi}\right]^{1/2} &\leq r && \text{(aplicando en ambos lados raíz cuadrada)} \\
0.5415 &\leq r
\end{aligned}
$$

Por lo tanto, el radio de la región circular de muestreo debe ser de al menos 0.5415 metros para poder hallar allí una planta o más, con probabilidad mayor o igual a 0.99.

■

EJERCICIO 2.6

Se está registrando la emisión de partículas radiactivas y se supone que es un proceso de Poisson con tasa 6 por minuto.

1. ¿Cuál es la probabilidad de que no haya registro de emisión de partículas en un período de 1 minuto?
2. ¿Cuál es la probabilidad de que en un período de 30 segundos ocurran al menos dos emisiones?
3. Si no hubo registro de emisión entre las 9:10 AM y las 9:12 AM, ¿cuál es la probabilidad de que ocurra una emisión entre las 10:10 AM y las 10:12 AM?
4. ¿Cuál es el período de tiempo para que la probabilidad que haya al menos una emisión sea mayor a 0.95?

PROPOSICIÓN 2.3: Si X tiene una distribución de Poisson con parámetro λ, $X \sim P(\lambda)$, entonces:

- $E(X) = \lambda$
- $V(X) = \lambda$
- $dt(X) = \sqrt{\lambda}$

Estos resultados también se pueden obtener de manera directa de las definiciones de media y varianza de una v.a. discreta.

Ejemplo 2.21

En base al Ejemplo 2.19, tenemos las siguientes v.a. $X_1 \sim P(5)$, $X_2 \sim P(10)$ y $X_{1/2} \sim P(2.5)$ entonces:

$$E(X_1) = V(X_1) = 5 \qquad dt(X_1) = 2.2361$$
$$E(X_2) = V(X_2) = 10 \qquad dt(X_2) = 3.1623$$
$$E(X_{1/2}) = V(X_{1/2}) = 2.5 \qquad dt(X_{1/2}) = 1.5811$$

■

Aproximación de Poisson a la binomial

Si $X \sim B(n, p)$, se puede demostrar que cuando n es grande y p pequeño, vale la siguiente aproximación:

$$f(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \cong e^{-\lambda} \frac{\lambda^k}{k!} \qquad k \in v_X \text{ y } \lambda = np$$

Es decir, $X \approx P(np)$. La notación $\approx$ significa que tiene aproximadamente esa distribución.

Esta aproximación es aceptable si $p \leq 0.05$ y $n \geq 20$.

Ejemplo 2.22

Se sabe que un peso muy bajo en el nacimiento, menor a 1500 gr, es una de las causas de mortalidad infantil. Se conoce que en determinada población, el porcentaje de niños con muy bajo peso al momento de nacer es de 1,2 %. Si consideramos 200 nacimientos en un hospital de esa población, ¿cuál es la probabilidad de que el número de recién nacidos con muy bajo peso en ese grupo sea mayor a 3?

Sea la v.a. $X =$"número de niños con muy bajo peso entre los 200 nacimientos de un hospital",

$X \sim B(200, 0.012)$ entonces:

$$P(X > 3) = 1 - P(X \leq 3) = 1 - \sum_{k=0}^{3} \binom{200}{k} 0.012^k (1 - 0.012)^{200-k} = 1 - 0.7795 = 0.2205$$

Como $p = 0.012 \leq 0.05$ y $n \geq 20$, se puede usar la aproximación de Poisson a la binomial y así facilitar las cuentas. Por lo tanto:

$$X \approx P(200 \times 0.012) \Leftrightarrow X \approx P(2.4)$$

Entonces:

$$P(X > 3) = 1 - P(X \leq 3) \cong 1 - e^{-2.4} \left[\frac{2.4^0}{0!} + \frac{2.4^1}{1!} + \frac{2.4^2}{2!} + \frac{2.4^3}{3!} \right] = 1 - 0.7787 = 0.2213$$

La siguiente tabla muestra que tan buena es la aproximación de sus frecuencias:

k	$B(200, 0.012)$	$P(2.4)$
0	0.0894105	0.0907179
1	0.2171917	0.2177231
2	0.2624766	0.2612677
3	0.2104063	0.2090142
4	0.1258605	0.1254085
5	0.0599238	0.0601960
6	0.0236541	0.0240784
7	0.0079622	0.0082554
8	0.0023330	0.0024766
9	0.0006045	0.0006604
10	0.0001402	0.0001585
11	0.0000294	0.0000345

■

Referencias

Cramer, H. (1968). *Elementos de la Teoría de Probabilidades y algunas de sus aplicaciones*. Madrid. Ed. Aguilar.

Devore Jay, L. (2001). *Probabilidad y Estadística para Ingeniería y Ciencias*. Ed. Books/Cole Publishing Company.

Feller, W. (1975). *Introducción a la Teoría de Probabilidades y sus Aplicaciones*. Ed. Limusa-Wiley S.A.

Maronna, R. (1995). *Probabilidad y Estadística Elementales para Estudiantes de Ciencias*. Buenos Aires. Ed. Exactas.

Mendenhall, W., Beaver, R. J. & Beaver, B. M. (2006). *Introducción a la Probabilidad y Estadística*. México. Cengage Learning Editores.

Meyer Paul, L. (1970). *Probabilidad y aplicaciones Estadísticas*. Addison-Wesley Iberoamericana.

Parzen, E. (1987). *Teoría Moderna de Probabilidades y sus Aplicaciones*. Ed. Limusa.

Ross, S. M. (1987). *Introduction to Probability and Statistics for Enginees and Scientists*. John Wiley & Sons.

Ross, S. M. (1997). *A first course in Probability*. New Jersey. Pearson Prentice Hall.

Walpole, R. E. & Myers, R. H. (2007). *Probabilidad y Estadística para Ingeniería y Ciencias*. México. Ediciones McGraw-Hill.

CAPÍTULO 3

Variables aleatorias continuas

En este capítulo estudiaremos variables aleatorias que pueden tomar valores en un intervalo de números reales.

Función de densidad de probabilidad

Definición:

Se dice que un v.a. X, que toma valores en un intervalo de números reales, es **continua** si existe una función f que cumple las siguientes condiciones:

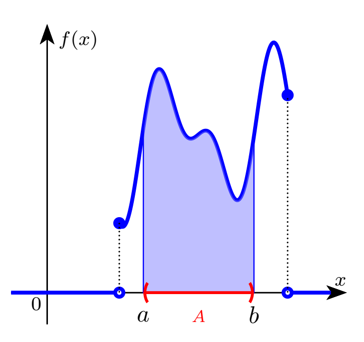
- $f(x) \geq 0$, para todo $x \in \mathbb{R}$,
- $\displaystyle\int_{-\infty}^{\infty} f(x)\, dx = 1$,
- $P(X \in A) = \displaystyle\int_{x \in A} f(x)\, dx$, para todo subconjunto $A \subseteq \mathbb{R}$.

La función f es llamada **función de densidad de probabilidad**, o simplemente *función de densidad* y la abreviaremos como fdp.

Observación:



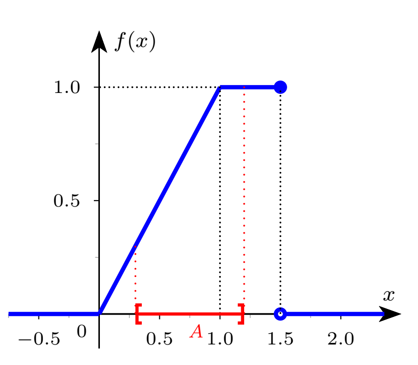
Para una v.a. continua X, la probabilidad de que tome valores en una región A incluida en $\mathbb{R}$, es igual al área bajo la curva densidad sobre esa región. Por ejemplo, si $A = (a, b)$, luego el área sombreada en la gráfica corresponde al valor $P(X \in A)$.



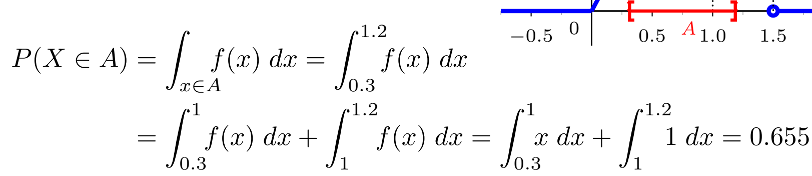
Ejemplo 3.1

Sea X una v.a. con función de densidad dada por:

$$f(x) = \begin{cases} x & si \quad 0 \leq x \leq 1 \\ 1 & si \quad 1 < x \leq 1.5 \\ 0 & cc \end{cases}$$



Luego si $A = [0.3, 1.2]$ podemos calcular:



$$\begin{aligned} P(X \in A) &= \int_{x \in A} f(x)\, dx = \int_{0.3}^{1.2} f(x)\, dx \\ &= \int_{0.3}^{1} f(x)\, dx + \int_{1}^{1.2} f(x)\, dx = \int_{0.3}^{1} x\, dx + \int_{1}^{1.2} 1\, dx = 0.655 \end{aligned}$$

■

Función de distribución o función de distribución acumulada

Igual que para una v.a. discreta, la función de distribución F de una v.a. X continua se define como:

$$F(x) = P(X \leq x) \text{ para todo } x \in \mathbb{R}.$$

y, en este caso, se calcula como:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)\, dy \tag{3.1}$$

de donde se deduce que la función de distribución de una v.a. continua, es una función continua.

Aplicando el Teorema Fundamental del Cálculo Integral en (3.1), se obtiene que la derivada de la función de distribución, en todos los puntos en los que la derivada existe, es la fdp:

$$f(x) = \frac{dF(x)}{dx} = F'(x).$$

La función F preserva las siguientes propiedades que vimos para el caso discreto, es decir:

- es una función no decreciente
- toma valores entre 0 y 1
- para todo $a, b \in \mathbb{R}$ tales que $a < b$ se cumple:

$$P(a < X \leq b) = F(b) - F(a)$$

Como ya se mencionó antes, la fda de una v.a. continua es una función continua.

Es importante resaltar que si X es una v.a. continua entonces:

$$P(X = a) = 0, \text{ para todo } a \in \mathbb{R}.$$

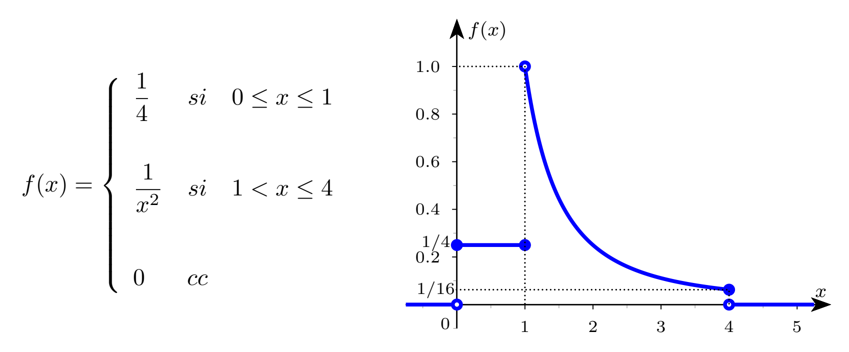
Luego, es evidente que para una v.a. continua y $a, b \in \mathbb{R}$:

$$P(a < X < b) = P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b)$$

Ejemplo 3.2

Sea X una v.a. con función de densidad dada por:

$$f(x) = \begin{cases} \frac{1}{4} & si \quad 0 \leq x \leq 1 \\ \frac{1}{x^2} & si \quad 1 < x \leq 4 \\ 0 & cc \end{cases}$$



Calculemos la fda de X, es decir, la expresión de la función $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt$. La función f no tiene la misma expresión en todo el eje real, ésto se muestra a continuación en un simple esquema:

0 1 4

$f(t)=0$ $f(t)=1/4$ $f(t)=1/t^2$ $f(t)=0$

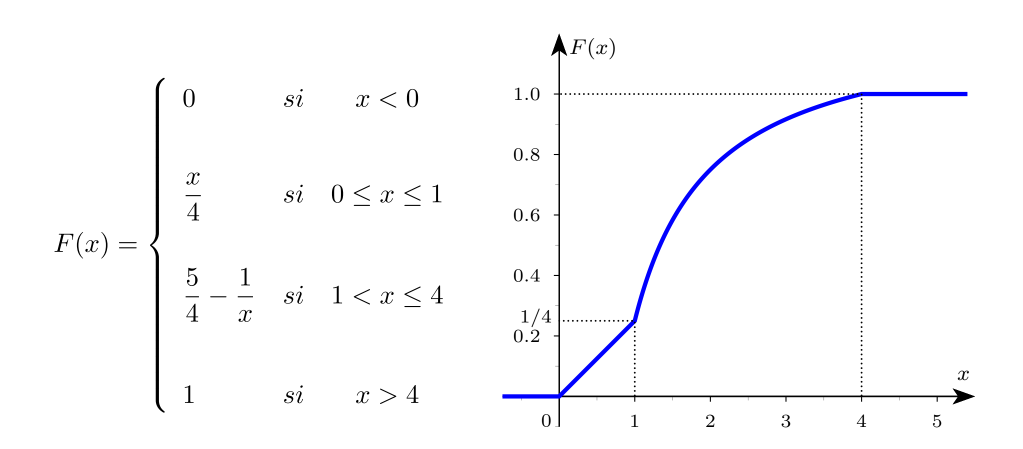
Para hallar la función F se procede de la siguiente manera:

- Si $x < 0$, $F(x) = \int_{-\infty}^{x} f(t)\, dt = \int_{-\infty}^{x} 0\, dt = 0$
- Si $0 \leq x \leq 1$, $F(x) = \int_{-\infty}^{x} f(t)\, dt = \int_{-\infty}^{0} f(t)\, dt + \int_{0}^{x} f(t)\, dt =$

$$\int_{-\infty}^{0} 0\, dt + \int_{0}^{x} \frac{1}{4}\, dt = \frac{x}{4}$$

- Si $1 < x \leq 4$, $F(x) = \displaystyle\int_{-\infty}^{x} f(t)\, dt = \int_{-\infty}^{0} f(t)\, dt + \int_{0}^{1} f(t)\, dt + \int_{1}^{x} f(t)\, dt =$
  $\displaystyle\int_{-\infty}^{0} 0\, dt + \int_{0}^{1} \frac{1}{4}\, dt + \int_{1}^{x} \frac{1}{t^2}\, dt = \frac{5}{4} - \frac{1}{x}$
- Si $x > 4$, $F(x) = \displaystyle\int_{-\infty}^{x} f(t)\, dt = \int_{-\infty}^{0} f(t)\, dt + \int_{0}^{1} f(t)\, dt + \int_{1}^{4} f(t)\, dt + \int_{4}^{x} f(t)\, dt =$
  $\displaystyle\int_{-\infty}^{0} 0\, dt + \int_{0}^{1} \frac{1}{4}\, dt + \int_{1}^{4} \frac{1}{t^2}\, dt + \int_{4}^{x} 0\, dt = 1$

Ésto se resume así:



$$F(x) = \begin{cases} 0 & si \quad x < 0 \\ \dfrac{x}{4} & si \quad 0 \leq x \leq 1 \\ \dfrac{5}{4} - \dfrac{1}{x} & si \quad 1 < x \leq 4 \\ 1 & si \quad x > 4 \end{cases}$$

Claramente, $F'(x) = f(x)$, para todo x salvo en 0, 1 y 4 (donde no es derivable).

■

EJERCICIO 3.1

1. Sea X una v.a. con función de distribución

$$F(x) = \begin{cases} 0 & si \quad x < -2 \\ \dfrac{1}{2} + \dfrac{3}{32}\left(4x - \dfrac{x^3}{3}\right) & si \quad -2 \leq x < 2 \\ 1 & si \quad x \geq 2 \end{cases}$$

   calcular y graficar la función de densidad.

2. Sea X una v.a. con función de densidad

$$f(x) = \begin{cases} 0 & si \quad x < 400 \\ 0.04e^{-0.04(x-400)} & si \quad x \geq 400 \end{cases}$$

   calcular y graficar la función de distribución acumulada.

Valor esperado o media

Definición:

Sea X una v.a. continua con fdp f. Entonces el **valor esperado** (llamado también *esperanza matemática*, *valor medio* o *media*) es

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx, \tag{3.2}$$

si se cumple que $\int_{-\infty}^{\infty} |x| f(x)\, dx < \infty$. Si esta integral diverge se dice que $E(X)$ no existe.

Ejemplo 3.3

Consideremos la v.a. X del Ejemplo 3.2 y calculemos su esperanza:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x)\, dx = \int_{-\infty}^{0} x f(x)\, dx + \int_{0}^{1} x f(x)\, dx + \int_{1}^{4} x f(x)\, dx + \int_{4}^{\infty} x f(x)\, dx \\ &= \int_{-\infty}^{0} x\, 0\, dx + \int_{0}^{1} x\, \frac{1}{4}\, dx + \int_{1}^{4} x\, \frac{1}{x^2}\, dx + \int_{4}^{\infty} x\, 0\, dx \\ &= \frac{1}{4} \int_{0}^{1} x\, dx + \int_{1}^{4} \frac{1}{x}\, dx = \frac{1}{4}\, \frac{x^2}{2}\bigg|_0^1 + \ln|x|\bigg|_1^4 = \frac{1}{8} + \ln 4 \cong 1.511 \end{aligned}$$

■

Valor esperado o media de una función de una v.a.

Si queremos calcular por definición la esperanza de una v.a. Y, que es función de una v.a. continua X, deberíamos calcular su fdp. Pero, como en el caso discreto, hay una propiedad que permite realizar el cálculo de la $E(Y)$ con la fdp de X.

PROPOSICIÓN 3.1: Sea X una v.a. continua con fdp f y sea $h : \mathbb{R} \to \mathbb{R}$ una función cualquiera, entonces $Y = h(X)$ es una v.a. cuya media se calcula como:

$$E(Y) = E\big(h(X)\big) = \int_{-\infty}^{\infty} h(x) f(x)\, dx,$$

si se cumple que $\int_{-\infty}^{\infty} |h(x)| f(x)\, dx < \infty$. Si esta integral diverge se dice que $E(Y)$ no existe.

La demostración de esta proposición está fuera del alcance de este curso.

Ejemplo 3.4

Consideremos la v.a. X del Ejemplo 3.1 y la función $h(X) = -2X^2$. Luego:

$$\begin{aligned}
E\big(h(X)\big) &= \int_{-\infty}^{\infty} h(x)f(x)\,dx = \int_{-\infty}^{\infty} \big(-2x^2\big)f(x)\,dx \\
&= \int_{-\infty}^{0}\big(-2x^2\big)f(x)\,dx + \int_{0}^{1}\big(-2x^2\big)f(x)\,dx + \int_{1}^{1.5}\big(-2x^2\big)f(x)\,dx + \int_{1.5}^{\infty}\big(-2x^2\big)f(x)\,dx \\
&= \int_{-\infty}^{0}\big(-2x^2\big)0\,dx + \int_{0}^{1}\big(-2x^2\big)\,x\,dx + \int_{1}^{1.5}\big(-2x^2\big)\,1\,dx + \int_{1.5}^{\infty}\big(-2x^2\big)0\,dx \\
&= -2\int_{0}^{1} x^3\,dx - 2\int_{1}^{1.5} x^2\,dx = -2.0833
\end{aligned}$$

■

Recordemos la siguiente propiedad vista para el caso discreto.

PROPIEDAD DE LINEALIDAD DE LA ESPERANZA: Sea X una v.a. tal que $E(X)$ existe y sean a y b dos reales, entonces

$$E(aX+b) = a\,E(X) + b$$

Demostración: Sea $f(x)$ la función de densidad de la v.a. X y consideremos la función $h(X) = aX + b$, $a, b \in \mathbb{R}$. Luego:

$$\begin{aligned}
E\big(aX+b\big) = E\big(h(X)\big) &= \int_{-\infty}^{\infty}(ax+b)f(x)\,dx && \big(\text{por Proposición 3.1}\big)\\
&= \int_{-\infty}^{\infty}\big(axf(x)+bf(x)\big)dx && \big(\text{distributiva}\big)\\
&= a\int_{-\infty}^{\infty}xf(x)\,dx + b\int_{-\infty}^{\infty}f(x)\,dx && \big(\text{por propiedades de integrales}\big)\\
&= aE(X)+b && \big(\text{por (3.2) y por definición}\big)
\end{aligned}$$

PROPIEDAD 3.1: Sea X una v.a. continua tal que $E(X)$ existe, entonces

$$E\big(X - E(X)\big) = 0$$

Demostración: Como $E(X)$ existe, la llamamos μ (que es un número). Consideremos la función $h(X) = X - \mu$. Notar que esta función es lineal en X, es decir, $h(X) = aX + b$, donde $a = 1$ y $b = -\mu$. Entonces, aplicando la Propiedad de Linealidad tenemos que:

$$E\big(h(X)\big) = E(X) + (-\mu) = \mu - \mu = 0$$

Varianza y desviación típica

Como ya dijimos en el caso discreto, la varianza de una v.a. da un idea de la dispersión de la distribución de la variable alrededor de su valor medio y se define como:

Definición:

Si X es una v.a. con media $E(X)$, definimos (si existe)

$$var(X) = E\Big[\big(X - E(X)\big)^2\Big]$$

y la desviación típica (o estándar)

$$dt(X) = \sqrt{var(X)}$$

Recordar que $dt(X)$ se expresa en las mismas unidades que la v.a. X.

También vale, para el caso continuo, la Propiedad 2.4: sea X una v.a. tal que existe su varianza y sean a y b dos constantes, entonces:

$$var(aX + b) = a^2 var(X) \tag{3.3}$$

Ejemplo 3.5

Considerar nuevamente la función de densidad del Ejemplo 3.1. Para poder calcular $var(X)$, primero debemos calcular $E(X)$:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x)\,dx = \int_{-\infty}^{0} x f(x)\,dx + \int_{0}^{1} x f(x)\,dx + \int_{1}^{1.5} x f(x)\,dx + \int_{1.5}^{\infty} x f(x)\,dx \\ &= \int_0^1 x^2\,dx + \int_1^{1.5} x\,dx = \frac{23}{24} \end{aligned}$$

Luego, por la Proposición 3.1 y considerando $h(X) = \big(X - E(X)\big)^2$:

$$\begin{aligned} var(X) &= E\Big[\big(X - E(X)\big)^2\Big] = E\left[\left(X - \frac{23}{24}\right)^2\right] = \int_{-\infty}^{\infty}\left(x - \frac{23}{24}\right)^2 f(x)\,dx \\ &= \int_{-\infty}^{\infty}\left[x^2 - 2x\frac{23}{24} + \left(\frac{23}{24}\right)^2\right] f(x)\,dx \\ &= \int_{-\infty}^{0}\left[x^2 - \frac{23}{12}x + \frac{529}{576}\right] 0\,dx + \int_0^1 \left[x^2 - \frac{23}{12}x + \frac{529}{576}\right] x\,dx \\ &+ \int_1^{1.5}\left[x^2 - \frac{23}{12}x + \frac{529}{576}\right] 1\,dx + \int_{1.5}^{\infty}\left[x^2 - \frac{23}{12}x + \frac{529}{576}\right] 0\,dx \\ &= \int_0^1 x^3\,dx - \frac{23}{12}\int_0^1 x^2\,dx + \frac{529}{576}\int_0^1 x\,dx + \int_1^{1.5} x^2\,dx - \frac{23}{12}\int_1^{1.5} x\,dx + \frac{529}{576}\int_1^{1.5} 1\,dx \\ &= \frac{1}{4} - \frac{23}{36} + \frac{529}{1152} + \frac{19}{24} - \frac{115}{96} + \frac{529}{1152} = \frac{71}{576} = 0.1233 \end{aligned}$$

EJERCICIO 3.2

La demanda semanal de gas propano (en miles de galones) de una distribuidora particular es una v.a. con fdp dada por:

$$f(x) = \begin{cases} 2\left(1 - \dfrac{1}{x^2}\right) & si \quad 1 \leq x \leq 2 \\ 0 & cc \end{cases}$$

Calcular la demanda semanal esperada y el desvío estándar.

Cálculo explícito de la varianza

De la misma manera que se realizó para v.a. discretas (en la Propiedad 3.3), para obtener explícitamente $var(X)$, desarrollamos el cuadrado en la definición y calculamos la media de la función de X usando la Proposición 3.1, obteniendo así:

$$var(X) = E(X^2) - \big(E(X)\big)^2$$

Demostración: Sea $f(x)$ la función de densidad de la v.a. X, $E(X) = \mu$ y $h(X) = (X - \mu)^2$. Entonces:

$$\begin{aligned} var(X) = E\big(h(X)\big) &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx && \big(\text{por Proposición 3.1}\big) \\ &= \int_{-\infty}^{\infty} (x^2 - 2x\mu + \mu^2) f(x)\, dx && \big(\text{desarrollando el cuadrado}\big) \\ &= \int_{-\infty}^{\infty} x^2 f(x)\, dx - \int_{-\infty}^{\infty} 2x\mu f(x)\, dx + \int_{-\infty}^{\infty} \mu^2 f(x)\, dx && \big(\text{distribuyendo}\big) \\ &= \int_{-\infty}^{\infty} x^2 f(x)\, dx - 2\mu \int_{-\infty}^{\infty} x f(x)\, dx + \mu^2 \int_{-\infty}^{\infty} f(x)\, dx \\ &= E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - 2\mu\mu + \mu^2 && \big(\text{por definición}\big) \\ &= E(X^2) - \mu^2 = E(X^2) - \big(E(X)\big)^2 \end{aligned}$$

Ejemplo 3.6

Sea X una v.a. con función de densidad dada en el Ejemplo 3.1. Veamos cuanto vale $var(3-2X)$. Podemos utilizar (3.3), donde $a = -2$ y $b = 3$, entonces

$$var(3 - 2X) = (-2)^2\, var(X) = 4 \times \frac{71}{576} = \frac{71}{144} \cong 0.4931$$

■

Si X tiene media μ y varianza σ^2 la "variable estandarizada"

$$Z = \frac{X - E(X)}{dt(X)} = \frac{X - \mu}{\sigma} \tag{3.4}$$

tiene media 0 y varianza 1, que se obtienen fácilmente de las propiedades de esperanza y varianza.

EJERCICIO 3.3

1. Demostrar que para la v.a. Z, definida en (3.4), vale $E(Z) = 0$ y $var(Z) = 1$.

2. En referencia al Ejercicio 3.2. Si hay 2500 galones en existencia y no se recibe nuevo suministro durante la semana. ¿Cuántos de los 2500 galones se espera que queden al fin de la semana? **Opcional**: ¿Cómo cambiarían los cálculos si la existencia al principio de la semana fuera de 1500 galones?

Otras medidas de resumen de una variable aleatoria

Definición:

Sea α un número entre 0 y 1, se define el **cuantil$-\boldsymbol{\alpha}$** de la v.a. $\boldsymbol{X}$ (discreta o continua) con función de distribución $\boldsymbol{F}$, como el número $\boldsymbol{x(\alpha)}$ tal que $\boldsymbol{F(x(\alpha)) = \alpha}$.
El cuantil$-\boldsymbol{\alpha}$ se suele llamar también percentil $\mathbf{100}\boldsymbol{\alpha}\,\%$.

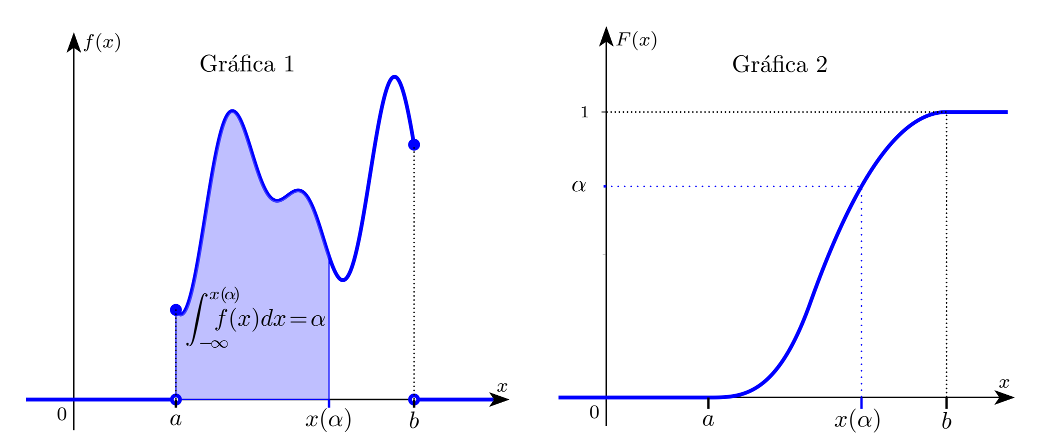
Aclaración

Los cuantiles para $\alpha = 0.25, 0.50$ y 0.75 se llaman **primer cuartil**, **mediana** o **segundo cuartil**, y **tercer cuartil** respectivamente.

Para v.a. continuas se calculan los cuantiles de la siguiente forma, para cualquier $0 < \alpha < 1$, el cuantil$-\alpha$, es el valor $x(\alpha)$, tal que:

$$F(x(\alpha)) = P(X \leq x(\alpha)) = \int_{-\infty}^{x(\alpha)} f(x)\, dx = \alpha$$

En los siguientes gráficos veremos como representar el cuantil α:



En la Gráfica 1, se ubicará $x(\alpha)$ en el eje x de modo que el área debajo de la curva de la función de densidad f desde $-\infty$ hasta $x(\alpha)$ sea igual a α. En la Gráfica 2 se ubicará el valor α sobre el eje y, trazando una recta paralela al eje x que pase por α, ésta cortará a la curva $F(x)$ en un punto cuya abscisa es $x(\alpha)$.

En particular, para $\alpha = 0.5$, la notación del cuantil-0.5 (mediana) es $\tilde{\mu}$

$$F(\tilde{\mu}) = \int_{-\infty}^{\tilde{\mu}} f(x)\, dx = 0.5$$

Ejemplo 3.7

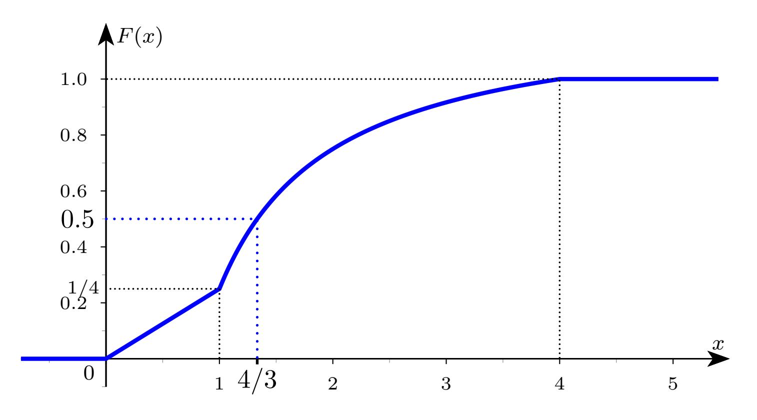
Sea X la v.a. del Ejemplo 3.2, busquemos su mediana. Recordemos que obtuvimos como f.d.a. de la v.a. X la siguiente expresión:

$$F(x) = \begin{cases} 0 & si \quad x < 0 \\ \dfrac{x}{4} & si \quad 0 \leq x \leq 1 \\ \dfrac{5}{4} - \dfrac{1}{x} & si \quad 1 < x \leq 4 \\ 1 & si \quad x > 4 \end{cases}$$

Como esta función es continua y no decreciente, hay que evaluarla en los valores donde cambia su definición para determinar el intervalo al que pertenece la mediana de la v.a. X. Es claro que en los intervalos $(-\infty, 0)$ y $(4, \infty)$ no va a estar el valor de $\tilde{\mu}$, pues $F(x) = 0$ para todo $x < 0$ y $F(x) = 1$ para todo $x > 4$. Al evaluar F en $x = 1$ se obtiene $F(1) = 1/4$ y como $1/4 < 1/2$, se puede asegurar que $\tilde{\mu} > 1$, es decir, $\tilde{\mu} \in (1, 4)$. Luego para los valores en ese intervalo planteamos:

$$F(\tilde{\mu}) = \frac{5}{4} - \frac{1}{\tilde{\mu}} = 0.5 \qquad \text{obteniendo que } \tilde{\mu} = \frac{4}{3}.$$

Veamos gráficamente su ubicación:



EJERCICIO 3.4

Para determinar el grado de inteligencia de un ratón, se mide el tiempo que tarda en recorrer un laberinto para encontrar la comida (estímulo). El tiempo (en segundos) que emplea un ratón es una v.a. Y con fdp dada por:

$$f(y) = \begin{cases} \dfrac{b}{y^2} & si \quad y \geq b \\ 0 & cc \end{cases}$$

donde $b > 0$. Calcular los cuartiles de la distribución. Graficar.

¿Es posible calcular su esperanza? Justificar.

Algunas variables aleatorias continuas

Distribución uniforme

Ejemplo 3.8

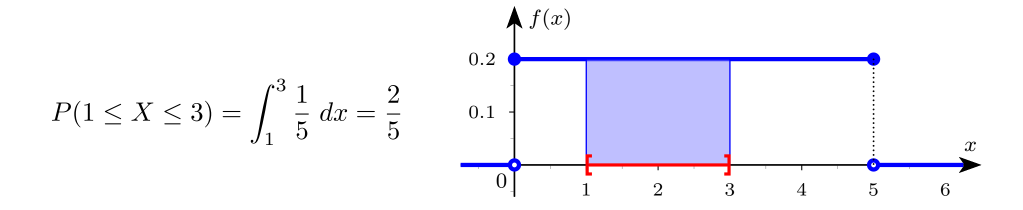
Supongamos que una persona toma un colectivo para ir al trabajo, que pasa exactamente cada 5 minutos. Si sale de su casa sin tener en cuenta la hora, el tiempo (medido en minutos) que tiene que esperar en la parada es una v.a. X, que puede tomar cualquier valor en el intervalo $[0, 5]$, la función de densidad para esta variable es:

$$f(x) = \begin{cases} 1/5 & si \quad 0 \leq x \leq 5 \\ 0 & cc \end{cases}$$

Es evidente que $f(x) \geq 0$ para todo x y que el área total bajo $f(x)$ es igual a 1.

Luego, la probabilidad de que tenga que esperar entre 1 y 3 minutos es:

$$P(1 \leq X \leq 3) = \int_1^3 \frac{1}{5}\, dx = \frac{2}{5}$$



Este ejemplo es el de una v.a. con distribución uniforme en el intervalo $[0, 5]$.

Definición:

Una v.a. tiene distribución **uniforme** en el intervalo $[a, b]$ $(a, b \in \mathbb{R}$ y $a < b)$, si su función de densidad está dada por:

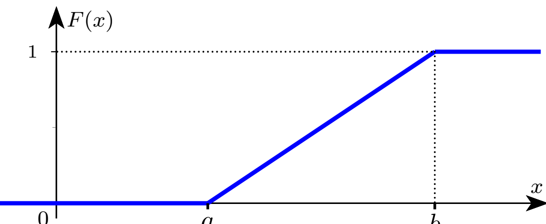
$$f(x) = \begin{cases} \dfrac{1}{b-a} & si \quad a \leq x \leq b \\ 0 & cc \end{cases}$$

Notación

Si la v.a. X tiene distribución uniforme en el intervalo $[a, b]$ lo denotamos como $X \sim U[a, b]$.

Si $X \sim U[a, b]$ su función de densidad f cumple: $f(x) \geq 0$ para todo $x \in \mathbb{R}$ y la $\int_{-\infty}^{\infty} f(x)\, dx = 1$.
Su función de distribución está dada por:

$$F(x) = \begin{cases} 0 & si \quad x < a \\ \dfrac{x-a}{b-a} & si \quad a \leq x \leq b \\ 1 & si \quad x > b \end{cases}$$



PROPOSICIÓN 3.2: Si $X \sim U[a, b]$ entonces:

- $E(X) = \dfrac{b+a}{2}$
- $var(X) = \dfrac{(a-b)^2}{12}$

Demostración: Calculemos primero la media de la v.a. X:

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx = \int_a^b x \frac{1}{b-a}\, dx = \frac{1}{b-a} \left.\frac{x^2}{2}\right|_a^b = \frac{1}{2(b-a)}(b^2 - a^2)$$
$$= \frac{1}{2\cancel{(b-a)}} \cancel{(b-a)} (b+a) = \frac{b+a}{2},$$

es decir, la media de una distribución uniforme es el punto medio del intervalo.

Ahora, para calcular la varianza, calculamos primero

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \int_a^b x^2 \frac{1}{b-a}\, dx = \frac{1}{b-a} \left.\frac{x^3}{3}\right|_a^b = \frac{1}{3(b-a)}(b^3 - a^3)$$
$$= \frac{1}{3\cancel{(b-a)}} \cancel{(b-a)}(a^2 + ab + b^2) = \frac{a^2 + ab + b^2}{3}.$$

Luego

$$var(X) = E(X^2) - (EX)^2 = \frac{a^2 + ab + b^2}{3} - \left(\frac{b+a}{2}\right)^2 = \frac{a^2 + ab + b^2}{3} - \frac{b^2 + 2ab + a^2}{4}$$
$$= \frac{4(a^2 + ab + b^2) - 3(b^2 + 2ab + a^2)}{12} = \frac{a^2 - 2ab + b^2}{12} = \frac{(a-b)^2}{12}$$

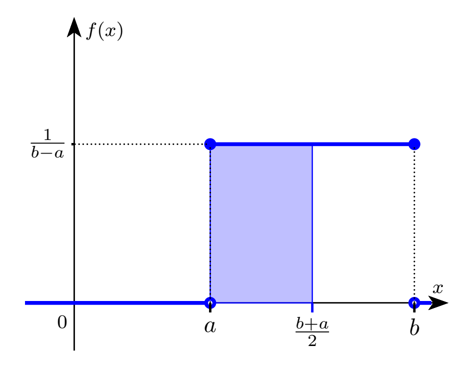
Si $X \sim U[a,b]$, para calcular su mediana, buscamos cuál es el valor $\tilde{\mu}$ para el cual $F(\tilde{\mu}) = 0.5$. Es claro que este valor se encuentra entre a y b, entonces planteamos:

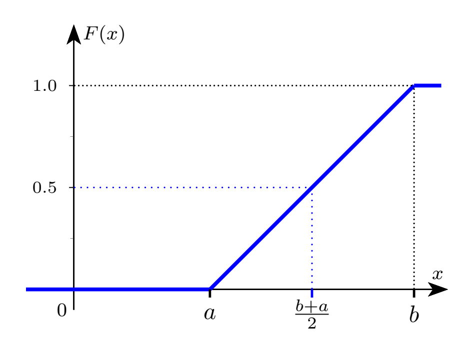
$$F(\tilde{\mu}) = \frac{\tilde{\mu} - a}{b - a} = 0.5$$

y despejando, obtenemos:

$$med(X) = \tilde{\mu} = \frac{b + a}{2}$$

En este caso la mediana y la media coinciden. Esto ocurre siempre que la distribución es simétrica. Gráficamente tenemos:





EJERCICIO 3.5

Si un paracaidista cae en un sitio aleatorio de la línea entre los puntos A y B.

1. Encuentre la probabilidad de que caiga más cerca de A que de B.
2. Calcule la probabilidad de que la distancia con respecto a A, sea más de 3 veces la distancia respecto de B.

Distribución exponencial

Definición:

Una v.a. X se dice que tiene distribución **exponencial** con parámetro λ, con $\lambda > 0$, si su función de densidad está dada por:

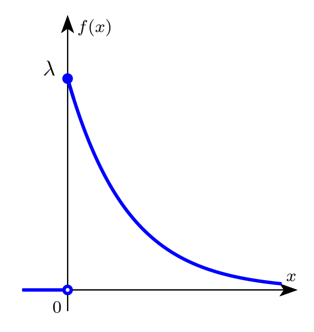
$$f(x) = \begin{cases} \lambda e^{-\lambda x} & si \quad x \geq 0 \\ 0 & si \quad x < 0 \end{cases}$$

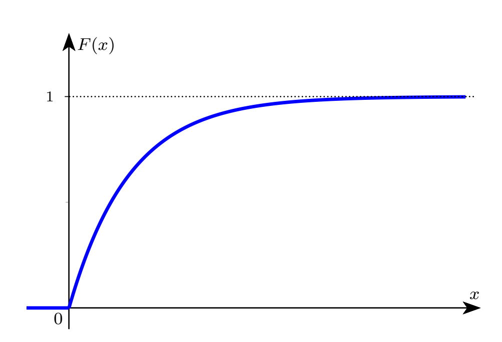
Notación

Si X tiene distribución exponencial con parámetro λ lo denotamos como $X \sim Exp(\lambda)$.

Si $X \sim Exp(\lambda)$ es fácil ver que su función de densidad cumple con $f(x) \geq 0$ para todo $x \in \mathbb{R}$, y la $\displaystyle\int_{-\infty}^{\infty} f(x)\, dx = 1$. Su función de distribución está dada por:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt = \begin{cases} 1 - e^{-\lambda x} & si \quad x \geq 0 \\ 0 & si \quad x < 0 \end{cases}$$





PROPOSICIÓN 3.3: Si $X \sim Exp(\lambda)$ entonces:

- $E(X) = 1/\lambda$
- $var(X) = 1/\lambda^2$

Demostración: Primero, por definición, tenemos que:

$$\begin{aligned} E(X) = \int_{-\infty}^{\infty} x f(x)\, dx = \int_{0}^{\infty} x\lambda e^{-\lambda x}\, dx = \lambda \int_{0}^{\infty} x e^{-\lambda x}\, dx \\ = \lambda \lim_{b\to\infty} \left[\frac{-xe^{-\lambda x}}{\lambda}\Big|_0^b + \int_0^b \frac{e^{-\lambda x}}{\lambda}\, dx \right] && \text{(integrando por partes)} \\ = \lambda \left[0 + \lim_{b\to\infty} \frac{-e^{-\lambda x}}{\lambda^2}\Big|_0^b \right] = \frac{1}{\lambda} \end{aligned}$$

Luego, por otro lugar, tenemos que:

$$\begin{aligned} E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \lambda \int_{0}^{\infty} x^2 e^{-\lambda x}\, dx \\ = \lambda \lim_{b\to\infty} \left[\frac{-x^2 e^{-\lambda x}}{\lambda}\Big|_0^b + 2\int_0^b \frac{x e^{-\lambda x}}{\lambda}\, dx \right] && \text{(integrando por partes)} \\ = \lambda \left[0 + \frac{2}{\lambda} \lim_{b\to\infty} \left(\int_0^b x e^{-\lambda x}\, dx \right) \right] \\ = 2 \lim_{b\to\infty} \left[\frac{-xe^{-\lambda x}}{\lambda}\Big|_0^b + \int_0^b \frac{e^{-\lambda x}}{\lambda}\, dx \right] && \text{(integrando por partes)} \\ = 2\,\frac{1}{\lambda^2} = \frac{2}{\lambda^2} \end{aligned}$$

Por último

$$var(X) = E(X^2) - \Big(E(X)\Big)^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

Si $X \sim Exp(\lambda)$, para calcular la mediana, buscamos el valor de $\tilde{\mu}$ tal que $F(\tilde{\mu}) = 0.5$. Sabemos

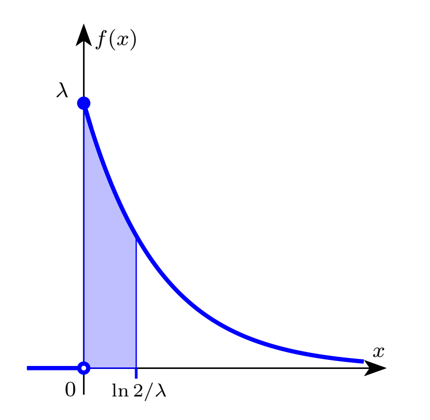
que $\tilde{\mu} \geq 0$, entonces planteamos:

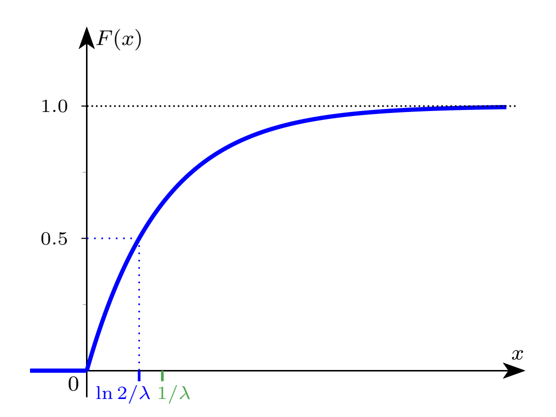
$$F(\tilde{\mu}) = 1 - e^{-\lambda\tilde{\mu}} = 0.5$$

Despejando, obtenemos que:

$$med(X) = \tilde{\mu} = \frac{\ln 2}{\lambda}$$

En este caso puede verse que la mediana es menor que la media $(1/\lambda)$.





PROPIEDAD DE AUSENCIA DE MEMORIA: Dada una v.a. $T \sim Exp(\lambda)$, si $t > 0$ y $s > 0$ se cumple que:

$$P(T > t + s | T > s) = P(T > t) \tag{3.5}$$

Demostración: Aplicando la definición de probabilidad condicional, tenemos que:

$$\begin{aligned}
P(T > t + s | T > s) &= \frac{P[(T > t + s) \cap (T > s)]}{P(T > s)} \\
&= \frac{P(T > t + s)}{P(T > s)} = \frac{1 - P(T \leq t + s)}{1 - P(T \leq s)} \\
&= \frac{1 - F(t + s)}{1 - F(s)} = \frac{1 - \left(1 - e^{-\lambda(t+s)}\right)}{1 - \left(1 - e^{-\lambda s}\right)} \\
&= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = \frac{e^{-\lambda t}\, e^{-\lambda s}}{e^{-\lambda s}} = e^{-\lambda t}
\end{aligned}$$

Ahora, si calculamos la probabilidad del lado derecho en (3.5) obtenemos que:

$$P(T > t) = 1 - P(T \leq t) = 1 - F(t) = 1 - \left(1 - e^{-\lambda t}\right) = e^{-\lambda t}$$

En consecuencia $P(T > t + s | T > s) = P(T > t)$.

Ejemplo 3.9

Sea X una v.a. con distribución exponencial con $\lambda = 0.2$. Sabiendo que $X > 2$, calculemos la probabilidad de que $X > 5$:

$$P(X > 5 \,|\, X > 2\,) = P(X > 3) = e^{-3\times 0.2} = 0.5488$$

■

Relación con un proceso de Poisson

Pensemos en un proceso temporal de Poisson de tasa c, y consideramos el tiempo T transcurrido desde el inicio del proceso hasta que se presenta el primer "evento" del mismo. Entonces, para cualquier $t > 0$, el evento $(T > t)$ significa que en el intervalo de tiempo $[0, t]$ no ocurre ninguno de los "eventos" del proceso. Si definimos X_t como el número de "eventos" que ocurren en el intervalo $[0, t]$, sabemos que $X_t \sim P(ct)$ y podemos decir:

$$P(T > t) = P(X_t = 0) = e^{-ct} \qquad \text{(recordar que } t > 0\text{)}$$

Entonces podemos determinar la función de distribución de esta v.a. T:

$$P(T \leq t) = 1 - P(T > t) = 1 - e^{-ct} \quad \text{si } t > 0$$

así hemos demostrado que $T \sim E(c)$.

También puede demostrarse que el tiempo transcurrido entre dos "eventos" sucesivos de un proceso de Poisson de tasa c, tiene distribución exponencial con parámetro $\lambda = c$.

Ejemplo 3.10

Suponga que se reciben llamadas en una línea telefónica de emergencias las 24 hs del día, según un proceso de Poisson con una tasa 0.5 llamadas por hora. ¿Cuál es la probabilidad de que transcurran más de dos horas entre dos llamadas sucesivas?

La v.a. T, el tiempo (medido en horas) entre dos llamados sucesivos, tiene distribución exponencial con $\lambda = 0.5$. Entonces

$$P(T > 2) = 1 - P(T \leq 2) = 1 - (1 - e^{-0.5\times 2}) = 0.3679$$

es la probabilidad pedida.

■

EJERCICIO 3.6

Sea X el tiempo (en días) entre dos fallas sucesivas de un equipo, X es una v.a. que tiene distri-

bución exponencial con $\lambda = 0.023$. Calcular:

1. La probabilidad de que el equipo funcione sin fallas durante al menos 42 días.
2. Si el equipo ha trabajado sin fallas durante 12 días, ¿cuál es la probabilidad de que en total funcione sin fallas más de 30 días?
3. ¿Cuál es la probabilidad de que haya más de un falla en 30 días?

La distribución normal

Definición:

Una v.a. X tiene distribución **normal** con parámetros μ (donde $\mu \in \mathbb{R}$) y σ^2 (donde $\sigma > 0$) si su función de densidad está dada por:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-(x-\mu)^2/(2\sigma^2)} \quad \text{para todo } x \in \mathbb{R}$$

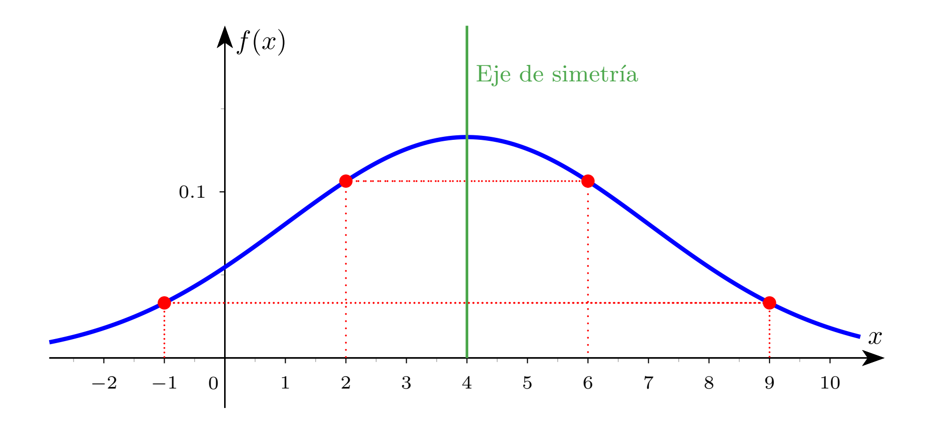
Notación

Si la v.a. X tiene distribución normal con parámetros μ y σ^2, se indica $X \sim N(\mu, \sigma^2)$.



Observación:

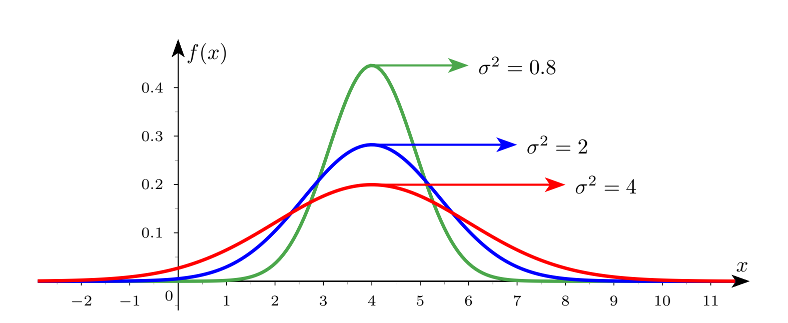
La gráfica de la función de densidad f de una v.a. $X \sim N(\mu, \sigma^2)$ es simétrica respecto de μ, pues $f(\mu + x) = f(\mu - x)$.





Observación:

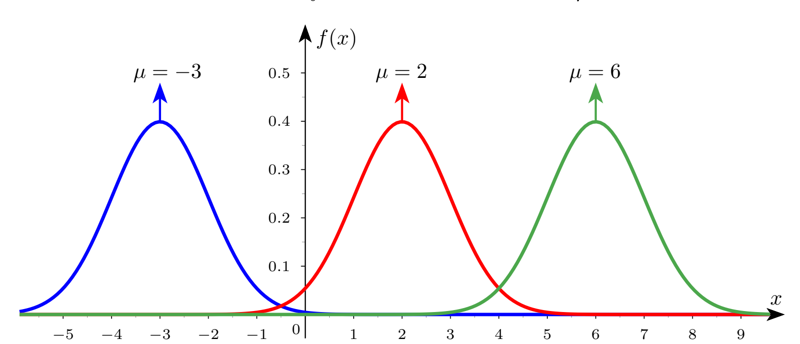
El siguiente gráfico corresponde a las funciones de densidad f de variables con distribución normal con $\mu = 4$ y diferentes valores de σ^2.





Observación:

El siguiente gráfico corresponde a las funciones de densidad f de variables con distribución normal con $\sigma^2 = 1$ y diferentes valores de μ.



PROPOSICIÓN 3.4: Si $X \sim N(\mu, \sigma^2)$ entonces:

- $E(X) = \mu$
- $var(X) = \sigma^2$
- $dt(X) = \sigma$

La demostración de esta proposición escapa los alcances de este libro.

En cuanto al cálculo de la función de distribución $F(x) = \int_{-\infty}^{x} f(t)\ dt$ no se puede encontrar una fórmula explícita de esta función, ya que la función $f(x)$ no tiene primitiva. Para calcular una probabilidad que involucre a una v.a. con distribución $N(\mu, \sigma^2)$ habría que hacer el cálculo utilizando métodos numéricos.

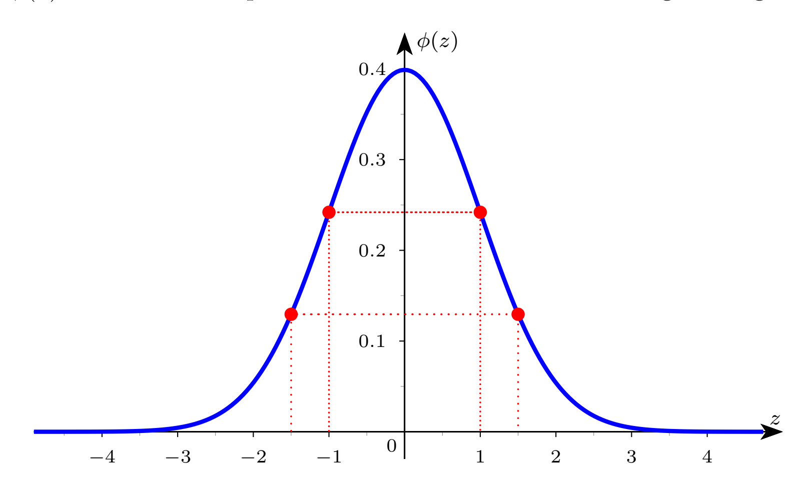
Veremos otro camino para resolver este problema en donde cobra un rol de gran importancia el caso de $\mu = 0$ y $\sigma = 1$.

Definición:

Una v.a. Z tiene distribución **normal típica** o *estándar* si su densidad está dada por:

$$\phi(z) = \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2} \quad \text{para } z \in \mathbb{R}$$

La función $\phi(z)$ es simétrica respecto de 0 como se muestra en la siguiente gráfica.



Notación

Si la v.a. Z tiene distribución normal típica, se indica $Z \sim N(0, 1)$.

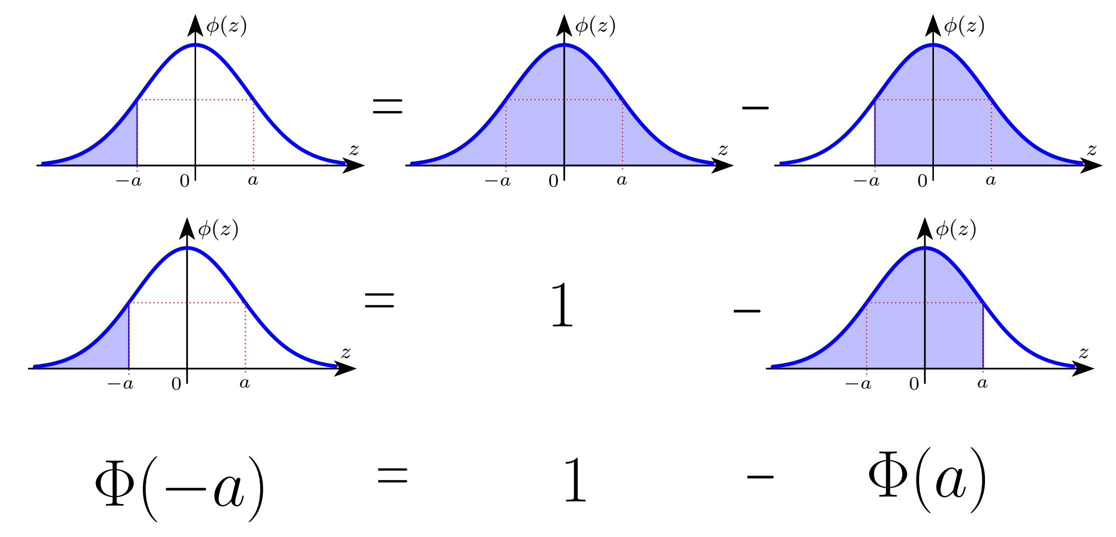
Si Z tiene esta distribución, entonces $E(Z) = 0$ y $var(Z) = dt(Z) = 1$, por la Proposición 3.4.

Llamamos $\Phi(z)$ a su función de distribución acumulada. Esta función se calcula por métodos numéricos y está tabulada para valores de z entre -3.49 y 3.49 (en el Apéndice B).

Por ser ϕ simétrica, la función Φ cumple:

$$\Phi(-a) = 1 - \Phi(a) \tag{3.6}$$

Gráficamente tenemos:



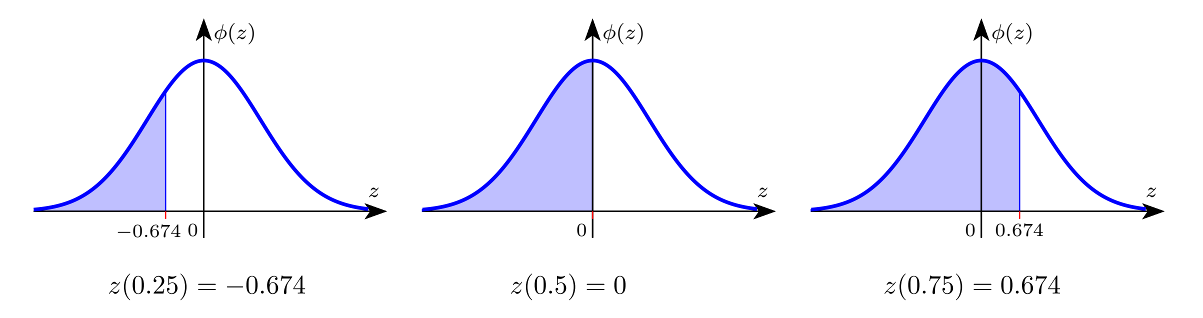
Se deduce de (3.6) que los cuantiles de la $N(0,1)$ cumplen

$$z(1-\alpha) = -z(\alpha)$$

pues $\Phi\big(-z(\alpha)\big) = 1 - \Phi\big(z(\alpha)\big) = 1 - \alpha$. En particular la mediana (o segundo cuartil)

$$med(Z) = \tilde{\mu} = z(0.5) = 0$$

A continuación se presentan en la gráfica de la función ϕ los 3 cuartiles de la v.a. $Z \sim N(0,1)$: primer, segundo y tercero.



$z(0.25) = -0.674$ $z(0.5) = 0$ $z(0.75) = 0.674$

PROPOSICIÓN 3.5: Si $X \sim N(\mu, \sigma^2)$, entonces para $a, b \in \mathbb{R}$ y cualquier $a \neq 0$, se verifica que:

$$Y = aX + b \sim N(a\mu + b, a^2\sigma^2)$$

Fácilmente se puede demostrar que $E(Y) = a\mu + b$ (por la Propiedad de Linealidad de la esperanza) y $Var(Y) = a^2\sigma^2$ (por (3.3)). Pero la demostración de que la v.a. Y tiene distribución normal está fuera del alcance de este libro.

COROLARIO 3.1: Si $X \sim N(\mu, \sigma^2)$ la variable estandarizada $Z = \dfrac{X - \mu}{\sigma}$ tiene distribución $N(0,1)$.

Demostración: Notemos que

$$Z = \frac{X - \mu}{\sigma} = \frac{1}{\sigma} X + \frac{-\mu}{\sigma}$$

Entonces, por la Proposición 3.5, definiendo $a = \frac{1}{\sigma}$ y $b = \frac{-\mu}{\sigma}$, tenemos que

$$a\mu + b = \frac{1}{\sigma}\mu + \frac{(-\mu)}{\sigma} = 0$$

y

$$a^2\sigma^2 = \left(\frac{1}{\sigma}\right)^2 \sigma^2 = 1$$

Por lo tanto, $Z \sim N(0,1)$.

Entonces, para calcular probabilidades correspondientes a una v.a. X con distribución $N(\mu, \sigma^2)$, se realiza la transformación $Z = \dfrac{X-\mu}{\sigma}$ y se trabaja con la distribución $N(0,1)$. Por lo tanto, si $a < b$:

$$
\begin{aligned}
P(a \leq X \leq b) &= P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\
&= P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \qquad \text{(por ser v.a. continua)} \\
&= P\left(\frac{a-\mu}{\sigma} < Z \leq \frac{b-\mu}{\sigma}\right) \\
&= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)
\end{aligned}
$$

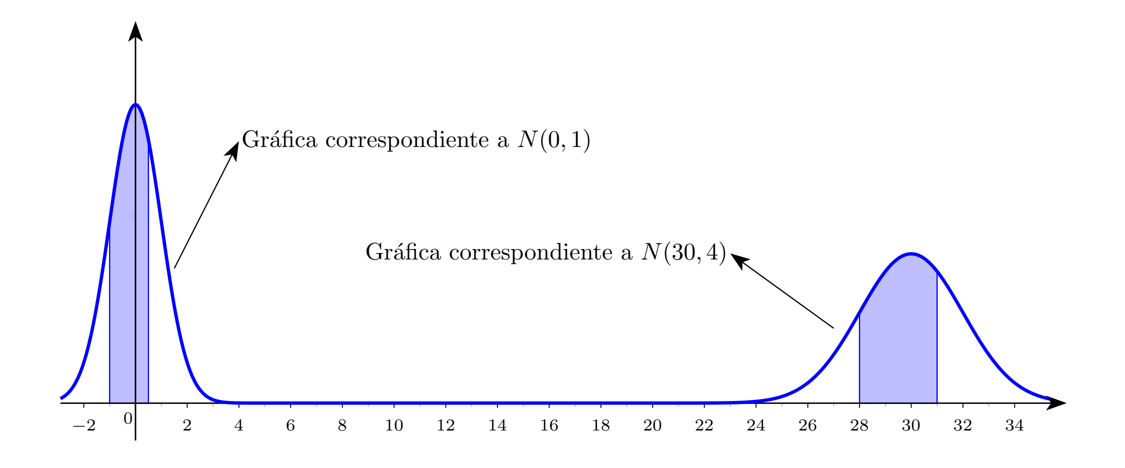
y de la misma forma:

$$
\begin{aligned}
F(x) &= P(X \leq x) \\
&= P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) \\
&= \Phi\left(\frac{x-\mu}{\sigma}\right) \qquad \text{para todo } x \in \mathbb{R}
\end{aligned}
$$

Ejemplo 3.11

Sea $X \sim N(30, 4)$, se desea calcular $P(28 < X < 31)$:

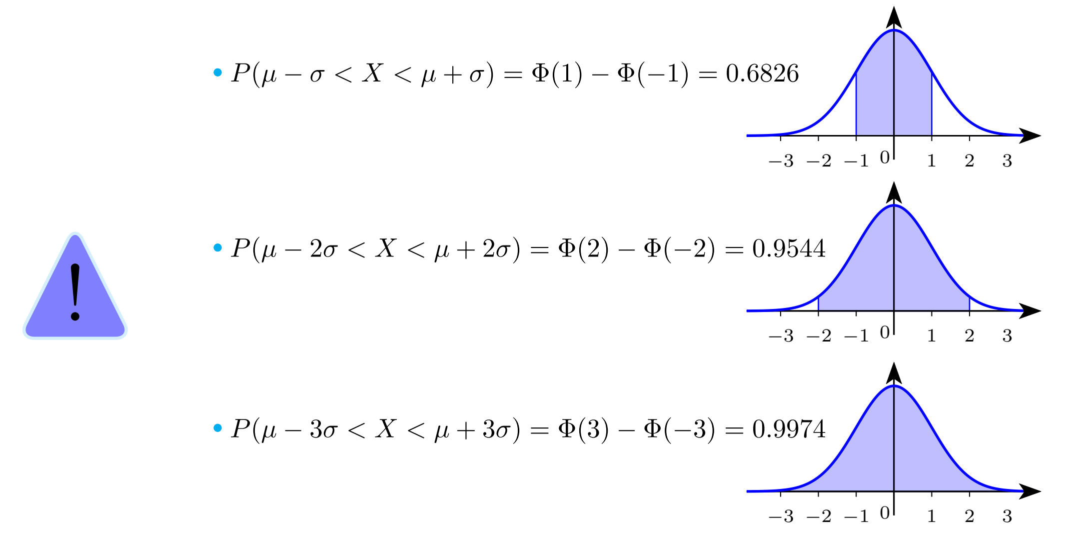
$$
\begin{aligned}
P(28 < X < 31) &= P\left(\frac{28-30}{2} < \frac{X-30}{2} < \frac{31-30}{2}\right) \\
&= P\left(-1 < \frac{X-30}{2} \leq 1/2\right) \\
&= P(-1 < Z \leq 1/2) \\
&= \Phi(0.5) - \Phi(-1) \\
&= 0.6915 - 0.1587 \\
&= 0.5328
\end{aligned}
$$



■

Observación:

Se puede ver que para cualquier variable aleatoria X con distribución $N(\mu, \sigma^2)$ se cumplen las siguientes igualdades:



- $P(\mu - \sigma < X < \mu + \sigma) = \Phi(1) - \Phi(-1) = 0.6826$
- $P(\mu - 2\sigma < X < \mu + 2\sigma) = \Phi(2) - \Phi(-2) = 0.9544$
- $P(\mu - 3\sigma < X < \mu + 3\sigma) = \Phi(3) - \Phi(-3) = 0.9974$

Estas igualdades caracterizan a la distribución normal.

En particular, aunque una v.a. normal puede tomar valores entre $-\infty$ e ∞, es "casi imposible" que tome valores que se alejen de μ en más de 3σ, ya que $P(|X - \mu| > 3\sigma) = 1 - 0.9974 = 0.0026$. Esta propiedad se utiliza, por ejemplo, en control de calidad.

Ejemplo 3.12

Una fábrica produce tornillos, las especificaciones indican que el diámetro de los mismos debe estar entre 1.19 y 1.21 pulgadas. Si el proceso de producción es tal que, el diámetro de los tornillos es una v.a. con distribución normal con media 1.196 y desviación estándar 0.005, ¿qué porcentaje de la producción no satisface las especificaciones?

Sea X la v.a. que mide el diámetro de los tornillos, $X \sim N(1.196, 0.005^2)$. Queremos calcular el porcentaje de la producción que no satisface las especificaciones, para ello calcularemos primero la probabilidad de que se cumplan las especificaciones:

$$P(1.19 < X < 1.21) = P\left(\frac{1.19 - 1.196}{0.005} < \frac{X - 1.196}{0.005} < \frac{1.21 - 1.196}{0.005}\right)$$
$$= P(-1.2 < Z < 2.8) = \Phi(2.8) - \Phi(-1.2) = 0.9974 - 0.1151 = 0.8823$$

Luego, por la propiedad de la probabilidad del complemento, la probabilidad de que no cumplan las especificaciones es $1 - 0.8823 = 0.1177$. Por lo tanto, el porcentaje de la producción que no satisface las especificaciones será 11.77 %.

■

Utilizando el Corolario 3.1, se prueba que los cuantiles de una variable X con distribución $N(\mu, \sigma^2)$ cumplen:

$$x(\alpha) = \mu + \sigma \times z(\alpha)$$

donde $z(\alpha)$ son los cuantiles de $N(0, 1)$.

En particular, la mediana de una v.a. $X \sim N(\mu, \sigma^2)$ es:

$$med(X) = x(0,5) = \mu$$

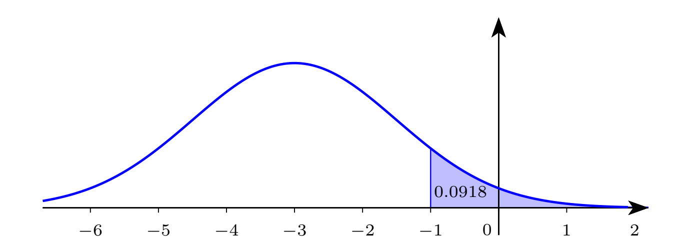
Ejemplo 3.13

Calcularemos los cuantiles 0.80 y 0.20 de una variable normal X con media 5 y desviación 2. De la Tabla: $\Phi(0.84) = 0.7995$ y $\Phi(0.85) = 0.8023$. Usando el valor más próximo $z(0.8) \cong 0.84$ y por lo tanto, $x(0.8) \cong 5 + 2 \times 0.84 = 6.68$. Para el otro cuantil usamos $z(0.2) = -z(0.8)$ y por lo tanto, $x(0.2) \cong 5 + 2 \times (-0.84) = 3.32$.

■

EJERCICIO 3.7

Hallar los parámetros μ y σ de la v.a. X normal cuya mediana vale -3 y la gráfica de su densidad es



Modelo de mediciones con error

Cuando se realiza cualquier medición, siempre se cometen errores. Los errores pueden ser sistemáticos, como en el caso de una balanza mal calibrada que siempre da un valor superior (o inferior) al verdadero, o pueden ser aleatorios. En muchas situaciones los errores sistemáticos pueden y deben ser eliminados, por ejemplo calibrando correctamente la balanza, en otras, cuando se conoce la magnitud del error, puede corregirse el resultado de la medición restando (o sumando, según sea el caso) el valor del error sistemático. Por otra parte los errores aleatorios nunca pueden ser completamente eliminados. Los errores que podemos tratar en estadística son los errores aleatorios. Debido a los **errores aleatorios** cuando repetimos una medición, en idénticas condiciones, los resultados de esas mediciones no son constantes, y varían alrededor de un valor medio.

Entonces, consideramos el resultado de una medición, como una variable aleatoria $X = \mu + \epsilon$, donde μ, si no hay error sistemático, es el verdadero valor de la magnitud que se está midiendo y ϵ es el error aleatorio, que tiene mediana 0 (esto indica que es igualmente probable que el error sea positivo o negativo).

Generalmente este tipo de error se modeliza con una distribución normal con media 0 y varianza σ^2, el valor de la varianza depende de la precisión del método de medición. Entonces, bajo este modelo, consideramos el resultado de una medición como una variable aleatoria: $X = \mu + \epsilon$, donde μ es el verdadero valor y ϵ tiene distribución $N(0, \sigma^2)$, lo que es equivalente: $X \sim N(\mu, \sigma^2)$.

Referencias

Cramer, H. (1968). *Elementos de la Teoría de Probabilidades y algunas de sus aplicaciones.* Madrid. Ed. Aguilar.

Devore Jay, L. (2001). *Probabilidad y Estadística para Ingeniería y Ciencias.* Ed. Books/Cole Publishing Company.

Feller, W. (1975). *Introducción a la Teoría de Probabilidades y sus Aplicaciones.* Ed. Limusa-Wiley S.A.

Maronna, R. (1995). *Probabilidad y Estadística Elementales para Estudiantes de Ciencias.* Buenos Aires. Ed. Exactas.

Mendenhall, W., Beaver, R. J. & Beaver, B. M. (2006). *Introducción a la Probabilidad y Estadística.* México. Cengage Learning Editores.

Meyer Paul, L. (1970). *Probabilidad y aplicaciones Estadísticas.* Addison-Wesley Iberoamericana.

Parzen, E. (1987). *Teoría Moderna de Probabilidades y sus Aplicaciones.* Ed. Limusa.

Ross, S. M. (1987). *Introduction to Probability and Statistics for Enginees and Scientists.* John Wiley & Sons.

Ross, S. M. (1997). *A first course in Probability.* New Jersey. Pearson Prentice Hall.

Walpole, R. E. & Myers, R. H. (2007). *Probabilidad y Estadística para Ingeniería y Ciencias.* México. Ediciones McGraw-Hill.

CAPÍTULO 4

Sumas de variables independientes y Teorema Central del Límite

En muchas situaciones experimentales se presentan más de una v.a. de interés. Aquí se describirán algunas combinaciones lineales de varias v.a., enfocándonos especialmente en el caso de v.a. independientes.

Esperanza y varianza de una combinación lineal de variables aleatorias

Dadas dos v.a. X e Y y dos constantes a, b reales, se verifica:

$$\begin{aligned} E(aX + bY) &= aE(X) + bE(Y) \\ var(aX + bY) &= a^2 var(X) + b^2 var(Y) + 2ab\, cov(X, Y) \end{aligned} \tag{4.1}$$

donde $cov(X, Y)$ es la covarianza entre X e Y, que es un parámetro que depende de la distribución conjunta de las v.a. Cuando las variables aleatorias son independientes $cov(X, Y) = 0$, entonces:

$$var(aX + bY) = a^2 var(X) + b^2 var(Y) \tag{4.2}$$



Observación:

Es importante recordar que la igualdad (4.2) es válida solamente cuando las variables aleatorias son *independientes*.

Ejemplo 4.1

Sea X una v.a. con media 2 y desvío 5. Sea Y otra v.a. con media -3 y varianza 4. Si X e Y son v.a. independientes, entonces:

$$E(2X - Y) = E\big(2X + (-1)Y\big) = 2E(X) + (-1)E(Y) = 2 \times 2 + (-1) \times (-3) = 7$$
$$var(2X - Y) = var\big(2X + (-1)Y\big) = 2^2 var(X) + (-1)^2 var(Y) = 4 \times 5^2 + 1 \times 4 = 104$$

■

Las igualdades (4.1) y (4.2) se generalizan para una combinación lineal de n variables aleatorias.

PROPOSICIÓN 4.1: Sean n v.a. X_1, X_2, ..., X_n y n constantes reales a_1, a_2, ..., a_n. Si Y es la combinación lineal $Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$, entonces la esperanza y la varianza de Y se calculan como:

$$E(Y) = E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i),$$
$$var(Y) = var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 Var(X_i) + 2\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j>i}}^{n} a_i a_j \, cov(X_i, X_j). \tag{4.3}$$

Si las v.a. X_i son independientes entre sí, tendremos que $cov(X_i, X_j) = 0$, para todo $i \neq j$, y la expresión (4.3) se reduce a:

$$var(Y) = var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \; var(X_i)$$

Destacamos especialmente el caso de v.a. independientes, porque en los próximos capítulos usaremos frecuentemente conjuntos de v.a. independientes.

Distribución de una combinación lineal de variables aleatorias normales independientes

La familia de distribuciones normales tiene una importante propiedad, dada en la siguiente proposición.

PROPOSICIÓN 4.2: Sean n v.a. independientes X_1, X_2, ..., X_n, donde cada X_i tiene distribución $N(\mu_i, \sigma_i^2)$ y sean n constantes reales, a_1, a_2, ..., a_n. Si la v.a. Y es la combinación lineal

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

entonces Y tiene distribución:

$$N\left(\sum_{i=1}^{n} a_i\, \mu_i, \sum_{i=1}^{n} a_i^2\, \sigma_i^2\right)$$

La demostración de esta propiedad está fuera del alcance de este curso.

Este resultado de la familia de distribuciones normales, sumamente importante, no vale en general para cualquier distribución.

Ejemplo 4.2

En una población adulta la distribución del peso es normal. Para las mujeres la media es 60 kg y el desvío 9 kg, para los hombres la media es 74 kg y el desvío 10 kg. Un hombre y dos mujeres entran a un ascensor cuya carga máxima es de 250 kg, ¿cuál es la probabilidad de que no se supere esa carga máxima?

Si definimos el peso de estas tres personas como las siguientes v.a.:

$$X = \text{“peso (en } kg\text{) de la primer mujer”},$$
$$Y = \text{“peso (en } kg\text{) de la segunda mujer” y}$$
$$Z = \text{“peso (en } kg\text{) del hombre”},$$

el peso total de estas tres personas es $W = X + Y + Z$. Por otro lado, como $X \sim N(60, 81)$, $Y \sim N(60, 81)$ y $Z \sim N(74, 100)$ son v.a. independientes entre sí, por la Proposición 4.2, tenemos que: $W \sim N(194, 262)$.

La probabilidad de que el peso total de estas personas no supere la carga máxima del ascensor es:

$$P(W \leq 250) = P\left(\frac{W - 194}{\sqrt{262}} \leq \frac{250 - 194}{\sqrt{262}}\right)$$
$$= \Phi(3.46) = 0.9997$$

■

EJERCICIO 4.1

1. Una compañía naviera maneja contenedores en tres diferentes tamaños: A_1 de 27 $pies^3$, A_2 de 125 $pies^3$ y A_3 de 512 $pies^3$. Sea X_i el número de contenedores de tipo A_i embarcados durante una semana dada, con $\mu_i = E(X_i)$ y $\sigma_i^2 = V(X_i)$, donde $i = 1, 2, 3$, dados en la

siguiente tabla:

Contenedor	μ	σ^2
A_1	200	10^2
A_2	250	12^2
A_3	300	8^2

- Suponiendo que X_1, X_2 y X_3 son v.a. independientes, calcule el valor esperado y la varianza del volumen total embarcado.

  (*Sugerencia*: Volumen $= 27 \times X_1 + 125 \times X_2 + 512 \times X_3$).

- ¿Serían sus cálculos necesariamente correctos si las v.a. X_i no fueran independientes? Explique.

2. Se supone que un comprimido contiene 500 mg de vitamina C, pero en realidad ese contenido es una variable aleatoria. Tenemos dos marcas de comprimidos, para la marca A el contenido de vitamina C tiene distribución $N(500, 22)$ y para la marca B el contenido de vitamina C es $N(500, 15)$. Si una persona durante una semana toma 2 comprimidos de la marca A y 5 de la marca B, ¿cuál es la probabilidad de que, en total, haya tomado menos de 3480 mg de vitamina C?

Propagación de errores

En muchos casos interesa conocer el valor de una medición indirecta, es decir, se está midiendo x, pero se desea conocer $h(x)$ y también interesa conocer cuál es el error en la medición indirecta $h(x)$, esto es lo que denominamos "propagación de errores". Por ejemplo, sabemos que la absorbancia a de una solución es el negativo del logaritmo (decimal) de su transmitancia t:

$$a = -\log t$$

Si la medición de la transmitancia es una v.a. X con desviación σ_X, entonces la absorbancia también es una v.a. Y:

$$Y = -\log X$$

Con estos elementos se desea calcular la desviación típica de Y, sabiendo que $E(X) = 0.501$ y su desvío estándar es $\sigma_X = 0.001$.

Para resolver este problema, lo planteamos de modo más general: sea una v.a. Y que es función de varias variables X_1, X_2, …, X_n independientes con medias μ_i y desviaciones típicas σ_i, $Y = h(X_1, X_2, \ldots, X_n)$, buscaremos una forma de aproximar su media y su varianza.

Recordemos previamente la aproximación mediante polinomios de Taylor de una función: la función $h(x)$ derivable en a, aproximada con el polinomio de Taylor de grado 1, en un entorno de a es:

$$h(x) \cong h(a) + h'(a)(x - a)$$

En caso de trabajar con una función de varias variables, diferenciable en $(a_1, a_2, \ldots, a_n)$, la correspondiente aproximación es:

$$\begin{aligned} h(x_1, x_2, \ldots, x_n) \cong h(a_1, a_2, \ldots, a_n) &+ \frac{\partial h}{\partial x_1}(a_1, a_2, \ldots, a_n)(x_1 - a_1) \\ &+ \frac{\partial h}{\partial x_2}(a_1, a_2, \ldots, a_n)(x_2 - a_2) + \cdots + \frac{\partial h}{\partial x_n}(a_1, a_2, \ldots, a_n)(x_n - a_n) \end{aligned}$$

Entonces trabajando con las v.a. X_1, X_2, ..., X_n el desarrollo de Taylor de grado 1, alrededor del punto $(\mu_1, \mu_2, \ldots, \mu_n)$, será:

$$\begin{aligned} Y = h(X_1, X_2, \ldots, X_n) \cong h(\mu_1, \mu_2, \ldots, \mu_n) &+ \frac{\partial h}{\partial X_1}(\mu_1, \mu_2, \ldots, \mu_n)(X_1 - \mu_1) \\ + \frac{\partial h}{\partial X_2}(\mu_1, \mu_2, \ldots, \mu_n)(X_2 - \mu_2) &+ \cdots + \frac{\partial h}{\partial X_n}(\mu_1, \mu_2, \ldots, \mu_n)(X_n - \mu_n) \end{aligned} \tag{4.4}$$

El lado derecho de esta ecuación es una función lineal de las v.a. X_1, X_2, ..., X_n. Si la distribución conjunta de X_1, X_2, ..., X_n está concentrada en una región en la cual la función h sea aproximadamente lineal, entonces (4.4) es una buena aproximación de Y. Por lo tanto, para aproximar los valores de $E(Y)$ y $dt(Y)$, podemos utilizar (4.4) y la Proposición 4.1:

$$\begin{aligned} E(Y) = E\big(h(X_1, X_2, \ldots, X_n)\big) \cong h(\mu_1, \mu_2, \ldots, \mu_n) &+ \frac{\partial h}{\partial X_1}(\mu_1, \mu_2, \ldots, \mu_n)E(X_1 - \mu_1) \\ + \frac{\partial h}{\partial X_2}(\mu_1, \mu_2, \ldots, \mu_n)E(X_2 - \mu_2) &+ \cdots + \frac{\partial h}{\partial X_n}(\mu_1, \mu_2, \ldots, \mu_n)E(X_n - \mu_n) \end{aligned}$$

Como $E(X_i - \mu_i) = 0$ para todo i, resulta:

$$E(h(X_1, X_2, \ldots, X_n)) \cong h(\mu_1, \mu_2, \ldots, \mu_n)$$

Del mismo modo, usando (4.4) y la Proposición 4.1, y recordando que las v.a. X_i son independientes:

$$\begin{aligned} var(Y) = var(h(X_1, X_2, \ldots, X_n)) &\cong \left(\frac{\partial h}{\partial X_1}(\mu_1, \mu_2, \ldots, \mu_n)\right)^2 var(X_1 - \mu_1) \\ + \left(\frac{\partial h}{\partial X_2}(\mu_1, \mu_2, \ldots, \mu_n)\right)^2 var(X_2 - \mu_2) &+ \cdots + \left(\frac{\partial h}{\partial X_n}(\mu_1, \mu_2, \ldots, \mu_n)\right)^2 var(X_n - \mu_n) \end{aligned}$$

Como $var(X_i - \mu_i) = var(X_i) = \sigma_i^2$ para todo i, resulta:

$$var(Y) \cong \left(\frac{\partial h}{\partial X_1}(\mu_1, \mu_2, ..., \mu_n)\right)^2 \sigma_1^2 + \left(\frac{\partial h}{\partial X_2}(\mu_1, \mu_2, ..., \mu_n)\right)^2 \sigma_2^2 + \cdots + \left(\frac{\partial h}{\partial X_n}(\mu_1, \mu_2, ..., \mu_n)\right)^2 \sigma_n^2$$

Por último, como

$$dt(Y) = dt(h(X_1, X_2, \ldots, X_n)) = \sqrt{var(h(X_1, X_2, \ldots, X_n))},$$

entonces

$$dt(Y) \cong \sqrt{\left(\frac{\partial h}{\partial X_1}(\mu_1, \mu_2, ..., \mu_n)\right)^2 \sigma_1^2 + \left(\frac{\partial h}{\partial X_2}(\mu_1, \mu_2, ..., \mu_n)\right)^2 \sigma_2^2 + \cdots + \left(\frac{\partial h}{\partial X_n}(\mu_1, \mu_2, ..., \mu_n)\right)^2 \sigma_n^2}$$

Ésto se resume en el siguiente resultado:

PROPOSICIÓN 4.3: Sean n v.a. independientes X_1, X_2, ..., X_n con $E(X_i) = \mu_i$ y $dt(X_i) = \sigma_i$. Si $Y = h(X_1, X_2, \ldots, X_n)$ para alguna función h, diferenciable en $(\mu_1, \mu_2, \ldots, \mu_n)$, entonces:

$$E(Y) \cong h(\mu_1, \mu_2, \ldots, \mu_n)$$

$$var(Y) \cong \sum_{i=1}^{n} \left(\frac{\partial h}{\partial X_i}(\mu_1, \mu_2, \ldots, \mu_n) \right)^2 \sigma_i^2$$

$$dt(Y) \cong \sqrt{\sum_{i=1}^{n} \left(\frac{\partial h}{\partial X_i}(\mu_1, \mu_2, \ldots, \mu_n) \right)^2 \sigma_i^2}$$

Continuando con el ejemplo de la absorbancia de una solución, tenemos que $h(x) = -\log x$, entonces para los $x > 0$, $h'(x) = -\log e/x = -0.434/x$ (con $e = 2.71828\ldots$), resulta:

$$E(Y) \cong -\log(0.501) = 0.30$$

y

$$dt(Y) \cong (0.434 \times 0.001)/0.501 = 0.0009$$

Ejemplo 4.3

Un péndulo simple se usa para medir la aceleración de la gravedad, usando la expresión:

$$T = 2\pi\sqrt{L/G}$$

donde T es el período de tiempo medido en segundos, L es la longitud del péndulo medida en metros y G es la aceleración de la gravedad. Las mediciones fueron hechas en un período de 1.24 seg y con una longitud de 0.38 m, donde las desviaciones fueron $\sigma_T = 0.02$ y $\sigma_L = 0.002$. ¿Cuál es valor resultante de la gravedad calculada y su desviación típica?

Consideramos las mediciones del período T y la longitud L como v.a., y tomamos como medias de las mismas los resultados de esta medición, es decir $\mu_T = 1.24seg$ y $\mu_L = 0.38m$, además sabemos que $\sigma_T = 0.02$ y $\sigma_L = 0.002$.

Para calcular la gravedad a partir de la expresión anterior, nos queda: $G = 4\pi^2 L/T^2$, entonces el valor calculado de la gravedad es una función de las v.a. longitud medida y período medido, es decir, $h(L, T) = 4\pi^2 L/T^2$. Luego

$$E(G) \cong h(0.38, 1.24) = 4\pi^2 0.38/1.24^2 = 9.7566\ m/seg^2$$

Para calcular la desviación típica, tenemos que:

$$\frac{\partial h}{\partial L}(L, T) = 4\pi^2/T^2 \quad \Rightarrow \quad \frac{\partial h}{\partial L}(\mu_L, \mu_T) = 4\pi^2/\mu_T^2 = 4\pi^2/1.24^2 = 25.6753$$

y

$$\frac{\partial h}{\partial T}(L,T) = -8\pi^2 L/T \quad \Rightarrow \quad \frac{\partial h}{\partial T}(\mu_L,\mu_T) = -8\pi^2\mu_L/\mu_T = -8\pi^2 0.38/1.24 = -24.1964$$

entonces:

$$\begin{aligned} var(G) &\cong 25.6753^2 \times 0.002^2 + (-24.1964)^2 \times 0.02^2 = 0.2368 \\ dt(G) &\cong \sqrt{0.2368} = 0.4866\ m/seg^2 \end{aligned}$$

■

EJERCICIO 4.2

Se desea calcular el volumen de un tanque cilíndrico, se mide el diámetro de la base (d) y la altura (h). Existen errores de medición, por ese motivo los resultados obtenidos se consideran variables aleatorias con $\sigma_d = 0.10\ m$ y $\sigma_h = 0.06\ m$. Consideramos los valores medidos $d = 4.50\ m$ y $h = 1.80\ m$ como las respectivas medias. Hallar la desviación estándar del volumen calculado.

Muestra aleatoria

Cuando se realizan mediciones repetidas (con error aleatorio) de una magnitud, los resultados de cada medición pueden ser representados por las v.a. $X_1, X_2, \ldots, X_n$, que se suponen independientes con la misma distribución, y por consiguiente con la misma media μ y la misma desviación típica σ. En ese caso, la media μ es la verdadera magnitud que deseamos medir y la desviación σ depende de la precisión del método de medición.

Cuando se eligen al azar n individuos de una población y se mide determinada característica de esos individuos (por ejemplo su nivel de hemoglobina), los resultados de esas mediciones también son variables aleatorias $X_1, X_2, \ldots, X_n$, independientes y con la misma distribución. En este caso μ y σ son la media y la desviación típica, respectivamente, de la distribución de esa característica en la población.

Definición:

Un conjunto de v.a. independientes y con la misma distribución se llama **muestra aleatoria** de esa distribución.

Notación

En general abreviaremos **muestra aleatoria** como m.a.

Definición:

Dada una m.a. $X_1, X_2, \ldots, X_n$, se define la media muestral o promedio muestral como:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$



Observación:

La media muestral es un caso particular de una combinación lineal donde $a_i = 1/n$ para todo i. Entonces, si $X_1, X_2, \ldots, X_n$ es una muestra aleatoria con $E(X_i) = \mu$ y $var(X_i) = \sigma^2$, aplicando la Proposición 4.1, la media muestral es una v.a. que cumple:

$$E(\overline{X}) = \mu, \quad var(\overline{X}) = \sigma^2/n \quad \text{y} \quad dt(\overline{X}) = \sigma/\sqrt{n} \tag{4.5}$$

A partir de (4.5) también se deduce que promediar varias mediciones aumenta la precisión, es decir, a mayor número de mediciones es mejor la precisión pues disminuye el desvío de $\overline{X}$. Si se tienen, por ejemplo, $n = 10$ mediciones y se desea duplicar la precisión o sea, reducir $dt(\overline{X})$ a la mitad hace falta tomar $n = 40$ (y no $n = 20$).

Ejemplo 4.4

Se van a realizar una serie de mediciones con un método que tiene un error de medición con $\sigma = 0.5$. ¿Cuántas mediciones se deben realizar si se desea que la desviación típica de la media muestral sea a lo sumo 0.2?

Los resultados de las n mediciones son una m.a. $X_1, X_2, \ldots, X_n$ y sabemos que $dt(\overline{X}) = 0.5/\sqrt{n} \leq 0.2$. De allí se deduce que $n \geq 2.5^2 = 6.25$, y por lo tanto es necesario hacer por lo menos 7 mediciones.

■

EJERCICIO 4.3

Considere el experimento de arrojar un dado equilibrado y observar el número de veces que sale el 2. Realice este experimento para poder calcular la media muestral y el desvío de la media muestral en cada uno de los siguientes casos:

- si se arroja el dado 5 veces
- si se arroja el dado 15 veces
- si se arroja el dado 30 veces

Distribución de $\overline{X}$

Como ya se mencionó, $\overline{X}$ es una variable aleatoria y su distribución depende de la distribución de las X_1, X_2, ..., X_n. En muchos casos conocer la forma de esta distribución no es simple, aunque siempre sabemos que la media de $\overline{X}$ es la misma que la de las X_i y la desviación típica de $\overline{X}$ es la misma de las X_i dividida por $\sqrt{n}$.

Sin embargo, hay una situación particular en que la distribución de $\overline{X}$ es sencilla, según la Proposición 4.2, cuando la m.a. X_1, X_2, ..., X_n tiene distribución normal $N(\mu, \sigma^2)$, $\overline{X}$ tiene distribución $N(\mu, \sigma^2/n)$.

Ahora bien, si la distribución de las X_1, X_2, ..., X_n no es normal, en general es difícil determinar la distribución de $\overline{X}$. Pero el siguiente resultado es de gran ayuda para poder aproximarla cuando podemos contar con una muestra suficientemente grande.

TEOREMA CENTRAL DEL LÍMITE (TCL): Sean n v.a., X_1, X_2, ..., X_n independientes entre sí y con la misma distribución, con media μ y desviación típica σ. Entonces si n es grande, $\overline{X}$ tiene aproximadamente una distribución normal con media μ y desviación típica $\sigma/\sqrt{n}$. Del mismo modo $\sum_{i=1}^{n} X_i$ también tiene distribución aproximadamente normal con media $n\mu$ y desviación típica $\sqrt{n}\sigma$.

Formalmente

$$P(\overline{X} \leq a) = P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{a - \mu}{\sigma/\sqrt{n}}\right) \cong \Phi\left(\frac{a - \mu}{\sigma/\sqrt{n}}\right)$$

y

$$P\left(\sum_{i=1}^{n} X_i \leq a\right) = P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \leq \frac{a - n\mu}{\sqrt{n}\sigma}\right) \cong \Phi\left(\frac{a - n\mu}{\sqrt{n}\sigma}\right)$$

Este teorema nos dice que cualquiera sea la distribución de la m.a., X_1, X_2, ..., X_n, podemos calcular (al menos en forma aproximada) probabilidades de eventos relacionados con $\sum_{i=1}^{n} X_i$, o con $\overline{X}$, siempre que n sea suficientemente grande. Como regla práctica, podemos decir que con $n \geq 30$ se consigue una aproximación aceptable.

Notación

Si $\overline{X}$ tiene aproximadamente una distribución normal con media μ y desviación típica $\sigma/\sqrt{n}$, lo abreviaremos, $\overline{X} \approx N(\mu, \sigma^2/n)$.

Si $\sum_{i=1}^{n} X_i$ también tiene distribución aproximadamente normal con media $n\mu$ y desviación típica $\sqrt{n}\sigma$, lo abreviaremos, $\sum_{i=1}^{n} X_i \approx N(n\mu, n\sigma^2)$.

Ejemplo 4.5

Supongamos que el consumo diario de calorías de una persona es una v.a. con media $\mu = 3000$ y desviación típica $\sigma = 350$. ¿Cuál es la probabilidad de que el consumo promedio de calorías diarias en el próximo año esté entre 2950 y 3050?

Los consumos diarios de cada uno de los días del próximo año, los podemos representar con las v.a. $X_1, X_2, \ldots, X_{365}$, que podemos suponer independientes y todas tienen la misma distribución desconocida con media 3000 y desviación típica 350. Nos interesa calcular la probabilidad de que el promedio de esas 365 v.a. esté entre 2950 y 3050, esto es:

$$
\begin{aligned}
P(2950 \le \overline{X} \le 3050) &= P\left(\frac{2950-3000}{350/\sqrt{365}} \le \frac{\overline{X}-3000}{350/\sqrt{365}} \le \frac{3050-3000}{350/\sqrt{365}}\right) && \text{(estandarizando)}\\
&\cong \Phi\left(\frac{3050-3000}{350/\sqrt{365}}\right) - \Phi\left(\frac{2950-3000}{350/\sqrt{365}}\right) && \text{(por TCL, pues } n \ge 30\text{)}\\
&= \Phi(2.73) - \Phi(-2.73) = 2\Phi(2.73) - 1 && \text{(por simetría de } \phi\text{)}\\
&= 2 \times 0.9968 - 1 = 0.9936 && \text{(por Tabla)}
\end{aligned}
$$

■

Aplicaciones particulares del Teorema Central del Límite

1. La distribución binomial, cuando n es grande, se puede aproximar a una normal con media np y varianza $np(1-p)$. Entonces si $X \sim B(n,p)$ se cumple:

$$P(X \le x) = P\left(\frac{X-np}{\sqrt{np(1-p)}} \le \frac{x-np}{\sqrt{np(1-p)}}\right) \cong \Phi\left(\frac{x-np}{\sqrt{np(1-p)}}\right)$$

Esta aproximación es aceptable cuando $np(1-p) > 5$.

Notación

Si $X \sim B(n,p)$, donde el valor de n es grande y $np(1-p) > 5$, entonces escribiremos $X \approx N\big(np, np(1-p)\big)$.

Ejemplo 4.6

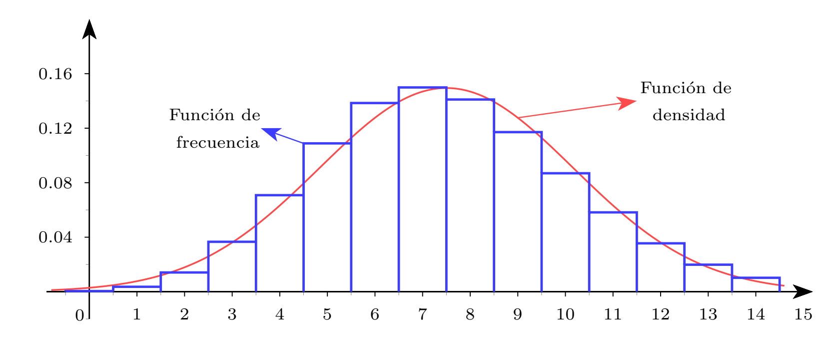
La primer tarea en un curso introductorio de programación por computadora implica correr un breve programa. Si la experiencia indica que 5 % de los estudiantes avanzados no cometen errores de programación, calcule la probabilidad de que en un grupo de 150 estudiantes que trabajan individualmente, menos de 3 alumnos no comentan errores.

Definimos la v.a. X como el número de alumnos, entre 150, que no cometen errores de programación, $X \sim B(150, 0.05)$, como n es grande y $np(1-p) = 150 \times 0.05 \times 0.95 = 7.125 > 5$

podemos aplicar el TCL:

$$P(X<3)=P(X\leq 2)=P\left(\frac{X-np}{\sqrt{np(1-p)}}\leq\frac{2-np}{\sqrt{np(1-p)}}\right)\cong\Phi(-2.06)=0.0197$$

Observar que el valor exacto es $P(X<3)=F(2)=0.01815406$. Gráficamente tenemos:



2. La distribución de Poisson, cuando λ es grande, se puede aproximar a una normal con media λ y varianza λ. Entonces si $X\sim P(\lambda)$ se cumple:

$$P(X\leq x)=P\left(\frac{X-\lambda}{\sqrt{\lambda}}\leq\frac{x-\lambda}{\sqrt{\lambda}}\right)\cong\Phi\left(\frac{x-\lambda}{\sqrt{\lambda}}\right)$$

Esta aproximación es aceptable cuando $\lambda>30$.

Notación

Si $X\sim P(\lambda)$, donde el valor de λ es grande, entonces escribiremos $X\approx N(\lambda,\lambda)$

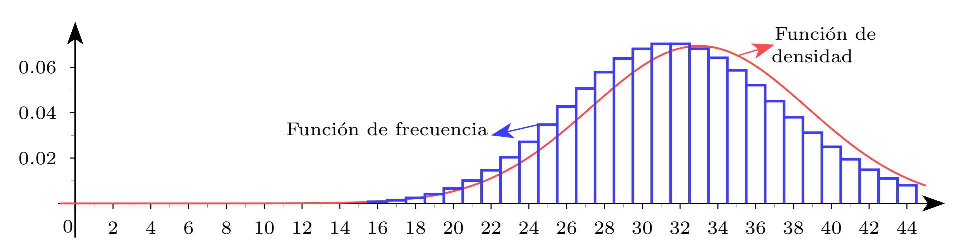
Ejemplo 4.7

El número de personas que entran, en un día, al puesto de periódicos a comprar una revista es una v.a. con distribución Poisson con $\lambda=32$. Encuentre la probabilidad de que a lo sumo 22 personas, en un día, entren al puesto de periódicos a comprar una revista.

Sea X el número de personas que entran a comprar una revista, entonces $X\sim P(32)$ y como λ es grande, podemos aplicar el TCL:

$$P(X\leq 22)=P\left(\frac{X-\lambda}{\sqrt{\lambda}}\leq\frac{22-\lambda}{\sqrt{\lambda}}\right)\cong\Phi(-1.77)=0.0384$$

Observar que el valor exacto es $P(X\leq 22)=F(22)=0.04062094$. Gráficamente tenemos:



EJERCICIO 4.4

1. El contenido de nicotina de un cigarrillo de una marca en particular es una v.a. con media 0.8 mg y desviación estándar 0.1 mg. Si un individuo fuma 5 paquetes de 20 de estos cigarrillos por semana, ¿cuál es la probabilidad de que la cantidad total de nicotina consumida en una semana sea por lo menos de 82 mg?

2. El número de colonias de bacterias de cierto tipo en una muestra de agua contaminada, tiene una distribución de Poisson con una media de 2 colonias por cm^3. ¿Cuál es la probabilidad de que en una muestra de 100 cm^3 haya más de 210 colonias?

3. Una línea aérea sabe que, en general, el 5 % de las personas que hacen sus reservaciones para cierto vuelo no se presentan. Si la aerolínea vende 160 boletos para un vuelo que tiene 155 lugares disponibles, ¿cuál es la probabilidad de que haya lugar disponible para todos los pasajeros que se presentan en el momento del vuelo?

Referencias

Cramer, H. (1968). *Elementos de la Teoría de Probabilidades y algunas de sus aplicaciones*. Madrid. Ed. Aguilar.

Devore Jay, L. (2001). *Probabilidad y Estadística para Ingeniería y Ciencias*. Ed. Books/Cole Publishing Company.

Feller, W. (1975). *Introducción a la Teoría de Probabilidades y sus Aplicaciones*. Ed. Limusa-Wiley S.A.

Maronna, R. (1995). *Probabilidad y Estadística Elementales para Estudiantes de Ciencias*. Buenos Aires. Ed. Exactas.

Mendenhall, W., Beaver, R. J. & Beaver, B. M. (2006). *Introducción a la Probabilidad y Estadística*. México. Cengage Learning Editores.

Meyer Paul, L. (1970). *Probabilidad y aplicaciones Estadísticas*. Addison-Wesley Iberoamericana.

Parzen, E. (1987). *Teoría Moderna de Probabilidades y sus Aplicaciones*. Ed. Limusa.

Ross, S. M. (1987). *Introduction to Probability and Statistics for Enginees and Scientists*. John Wiley & Sons.

Ross, S. M. (1997). *A first course in Probability*. New Jersey. Pearson Prentice Hall.

Walpole, R. E. & Myers, R. H. (2007). *Probabilidad y Estadística para Ingeniería y Ciencias*. México. Ediciones McGraw-Hill.

CAPÍTULO 5

Estimación

Introducción

El objetivo de la Estadística es hacer inferencias con respecto a una población, basándose en la información contenida en una muestra. Generalmente las distribuciones poblacionales dependen de ciertos parámetros. Por ejemplo en los Capítulos 2 y 3, estudiamos algunas distribuciones poblacionales discretas, como Binomial y Poisson, o continuas, como Exponencial, Uniforme y Normal. Cada una de ellas es descrita mediante uno o más parámetros, en la Binomial el parámetro p representa la probabilidad de éxito, en la Normal el parámetro μ es la esperanza o media de la población, mientras σ^2 es la varianza, y nos da una idea de la dispersión de los valores poblacionales respecto de μ, etc. Conociendo el valor de esos parámetros aprendimos a calcular cuán probables podrían ser ciertos resultados en la población. Sin embargo, en una situación real, esos parámetros poblacionales son desconocidos.

En este capítulo veremos cómo utilizar los valores de una muestra para obtener información sobre el valor de un parámetro poblacional. Veremos algunas propiedades de las funciones que suelen utilizarse para calcular estimaciones puntuales de los parámetros desconocidos y estudiaremos procedimientos para obtener estimaciones mediante intervalos de confianza.

Estimación puntual

El objetivo de la estimación puntual es, usando los datos muestrales, seleccionar un solo número que sea, en algún sentido, una buena presunción del valor del parámetro que se desea conocer.

Supongamos, por ejemplo, que el parámetro de interés es el valor medio de duración de baterías de cierto tipo para una calculadora. Una muestra aleatoria de 5 baterías podría dar duraciones observadas, en horas, de: $x_1 = 5$, $x_2 = 6.4$, $x_3 = 5.9$, $x_4 = 5.3$ y $x_5 = 5.7$. El valor calculado de la duración media muestral es $\overline{x} = 5.66$, es razonable considerar el valor 5.66 como una buena presunción del verdadero valor de la media poblacional.

Al analizar conceptos generales y métodos de inferencia, usaremos la letra griega θ para indicar el parámetro de interés.

Veremos algunas definiciones básicas.

Primero recordemos que una **muestra aleatoria** es un conjunto de v.a. $X_1, X_2, \ldots, X_n$ independientes y todas con la misma distribución. Los valores observados de esa m.a. son números $x_1, x_2, \ldots, x_n$.

Definición:

Sean $X_1, X_2, \ldots, X_n$ una m.a., llamamos **estadístico** a cualquier función de la m.a. Entonces, un estadístico es también una v.a.

Definición:

Si $X_1, X_2, \ldots, X_n$ es una m.a. de una distribución que depende de un parámetro θ, un **estimador puntual** de θ es un estadístico $h(X_1, X_2, \ldots, X_n)$, que se utiliza para estimar el valor desconocido de ese parámetro, de modo que, un **estimador** es una v.a. Cuando esa función se aplica a los valores observados de la m.a., $h(x_1, x_2, \ldots, x_n)$, constituye una **estimación puntual**, que es un número.

Notación

En general al estimador de θ lo denotamos $\widehat{\theta}$, también a su estimación.

Ejemplo 5.1

En el ejemplo anterior, el estimador puntual de la media verdadera de la duración de las baterías es $\overline{X} = (X_1 + X_2 + X_3 + X_4 + X_5)/5$, y el valor calculado para los valores muestrales, 5.66, es la estimación puntual.

■

Algunos estimadores básicos

1. Si tenemos una m.a. $X_1, X_2, \ldots, X_n$ de cualquier distribución que sabemos que tiene media μ, el estimador usual para este parámetro es la media muestral $\overline{X}$, que ya definimos anteriormente.

2. Si tenemos una m.a. $X_1, X_2, \ldots, X_n$, de una distribución continua desconocida, sabemos que para cualquier distribución existe la mediana $\widetilde{\mu}$. Vamos a ver cómo estimar $\widetilde{\mu}$.

Definición:

Dados los valores observados $x_1, x_2, \ldots, x_n$ de una m.a. Llamamos $x_{(i)}$ a los x_i ordenados:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

La **mediana muestral**, que denotaremos como $med(x_1, x_2, \ldots, x_n) = \widetilde{x}$, es el valor que verifica que la mitad de las observaciones son menores a $\widetilde{x}$ y la mitad de las observaciones son mayores a $\widetilde{x}$. La mediana muestral es una estimación de la mediana poblacional $\widetilde{\mu}$.

En la práctica, la mediana se calcula de la siguiente manera:

- si n es impar, entonces $\widetilde{x} = x_{(m)}$ con $m = (n+1)/2$
- si n es par, entonces $\widetilde{x} = \dfrac{x_{(m)} + x_{(m+1)}}{2}$ con $m = n/2$

Ejemplo 5.2

Consideremos las siguientes 20 observaciones, cada una representa la duración (en horas) de un cierto tipo de lámpara incandescente:

1088	666	1016	964	1058	612	1003	898	1197	1022
744	1135	623	1085	970	1201	983	1029	883	1122

Para calcular la mediana debemos ordenar las observaciones de menor a mayor:

612	623	666	744	883	898	964	970	983	1003
1016	1022	1029	1058	1085	1088	1122	1135	1197	1201

Como en este caso tenemos 20 observaciones, la mediana será el promedio de las dos centrales:

$$\widetilde{x} = \frac{x_{(10)} + x_{(11)}}{2} = \frac{1003 + 1016}{2} = 1009.5$$

■

3. Si tenemos una m.a. $X_1, X_2, \ldots, X_n$ de cualquier distribución que tiene media μ y varianza σ^2, el estimador usual para la varianza es la **varianza muestral**, que llamaremos S^2:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Ejemplo 5.3

Recordemos el ejemplo de las 5 baterías, donde las duraciones observadas, en horas, son: $x_1 = 5$, $x_2 = 6.4$, $x_3 = 5.9$, $x_4 = 5.3$ y $x_5 = 5.7$. Si calculamos la varianza muestral, se tiene que $S^2 = 0.293$.

■

4. Si $X \sim B(n, p)$ un posible estimador de p es $\widehat{p} = X/n$.

Ejemplo 5.4

Se desea evaluar la efectividad de un nuevo medicamento contra una enfermedad. Se administró el medicamento a 150 personas que padecían dicha enfermedad y se observó que 114 personas se recuperaron a los 3 días. Estime la proporción de individuos que se recuperan dentro de los tres días de administrado el medicamento.

En este ejemplo tenemos, $X =$"número de individuos que se recuperan dentro de los tres días de administrado el medicamento, entre 150" donde $X \sim B(150, p)$. Queremos estimar p, para ello tenemos que $x = 114$ y $n = 150$, entonces $\widehat{p} = \frac{114}{150} = 0.76$.

■

Algunas propiedades

Si queremos estimar un parámetro, no parece razonable elegir cualquier función de la muestra. En general pediremos que el estimador tenga algunas propiedades.

Definición:

Si $\widehat{\theta}$ es un estimador de θ, se llama **sesgo** del estimador a la diferencia $E(\widehat{\theta}) - \theta$.

Se dice que el estimador $\widehat{\theta}$ es **insesgado**, si $E(\widehat{\theta}) = \theta$, para cualquier valor de θ.

Ejemplo 5.5

Dada una m.a. de una distribución con media μ, ya vimos que $E(\overline{X}) = \mu$, luego $\overline{X}$ es un estimador insesgado de la media poblacional μ.

■

Ejemplo 5.6

Dada una m.a. de una distribución con media μ y varianza σ^2, la varianza muestral es un estimador insesgado de la varianza poblacional σ^2, esto significa que $E(S^2) = \sigma^2$.

Para probar esta propiedad, desarrollamos el cuadrado:

$$\sum_{i=1}^{n}(X_i-\overline{X})^2 = \sum_{i=1}^{n}\left(X_i^2-2X_i\overline{X}+\overline{X}^2\right) = \sum_{i=1}^{n}X_i^2 - 2\overline{X}\sum_{i=1}^{n}X_i + n\overline{X}^2$$
$$= \sum_{i=1}^{n}X_i^2 - 2\overline{X}n\overline{X} + n\overline{X}^2 = \sum_{i=1}^{n}X_i^2 - 2n\overline{X}^2 + n\overline{X}^2 \qquad \Big(\text{ya que } \sum_{i=1}^{n}X_i = n\overline{X}\Big)$$
$$= \sum_{i=1}^{n}X_i^2 - n\overline{X}^2$$

Luego, aplicando las propiedades de la esperanza:

$$E(S^2) = \frac{1}{n-1}\left(\sum_{i=1}^{n}E\left(X_i^2\right) - nE\left(\overline{X}^2\right)\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n}\left(\sigma^2+\mu^2\right) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right) \tag{5.1}$$
$$= \frac{1}{n-1}\left(n\left(\sigma^2+\mu^2\right)-\sigma^2-n\mu^2\right) = \frac{1}{n-1}\left(n\sigma^2+\cancel{n\mu^2}-\sigma^2-\cancel{n\mu^2}\right)$$
$$= \frac{1}{\cancel{n-1}}\left(\sigma^2\cancel{(n-1)}\right) = \sigma^2$$

En (5.1) se utilizo para el cálculo de $E(X_i^2)$ y $E(\overline{X}^2)$ que para cualquier v.a. Y tenemos que $E(Y^2) = var(Y) + \big(E(Y)\big)^2$ (por la Propiedad 2.3).

■

Ejemplo 5.7

Si $X \sim B(n,p)$, $\widehat{p} = X/n$ es un estimador insesgado para p, ya que:

$$E(\widehat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}\,E(X) = \frac{1}{\cancel{n}}\,\cancel{n}p = p$$

■

Ejemplo 5.8

Supongamos que el tiempo de espera para tomar un colectivo tiene una distribución uniforme en el intervalo de tiempo $[0,\theta]$ donde θ es desconocido. Se desea estimar al parámetro θ en base a los siguientes tiempos que tiene que esperar una persona durante 10 días:

4.5 6.3 3.1 1.1 8.9 2.4 0.6 7.3 5.7 9.2

Estos son los valores observados de una muestra aleatoria $X_1, X_2, \ldots, X_n$ de una distribución $U[0,\theta]$.

Como el valor de θ es el máximo tiempo de espera posible, parece razonable elegir como estimador de θ, al máximo de los tiempos de espera de la muestra, esto se escribe:

$$\widehat{\theta} = \text{máx}(X_1, X_2, \ldots, X_n).$$

Se puede demostrar (pero escapa al alcance de este curso) que:

$$E(\widehat{\theta}) = \frac{n}{n+1}\,\theta, \tag{5.2}$$

esto significa que, $\widehat{\theta}$ no es un estimador insesgado para θ, sin embargo, a partir del mismo, se puede obtener un estimador insesgado. Definimos a este nuevo estimador como:

$$\widehat{\theta}_1 = \frac{n+1}{n}\,\text{máx}(X_1, X_2, \ldots, X_n),$$

este estimador es insesgado, ya que:

$$\begin{aligned} E(\widehat{\theta}_1) &= E\left(\frac{n+1}{n}\,\text{máx}(X_1, X_2, \ldots, X_n)\right) = \frac{n+1}{n} E\big(\text{máx}(X_1, X_2, \ldots, X_n)\big) \\ &= \frac{n+1}{n}\, E(\,\widehat{\theta}\,) \overset{(5.2)}{=} \frac{\cancel{n+1}}{\cancel{n}}\,\frac{\cancel{n}}{\cancel{n+1}}\,\theta = \theta \end{aligned}$$

Por otro lado, recordando que la esperanza de una v.a. con distribución uniforme es el punto medio del intervalo, en este caso $\theta/2$, se podría definir otro estimador para θ como:

$$\widehat{\theta}_2 = 2\overline{X}$$

Este estimador también es insesgado, ya que:

$$E(\widehat{\theta}_2) = E(2\overline{X}) = 2E(\,\overline{X}\,) = 2\theta/2 = \theta$$

Luego, los valores de la estimación obtenida para el primer estimador, $\widehat{\theta}_1$, es $\frac{11}{10}\,\text{máx}(x_1, x_2, \ldots, x_n) = \frac{11}{10} \times 9.2 = 10.12$. Y para el segundo $\widehat{\theta}_2$ es $2\,\overline{x} = 2 \times 4.91 = 9.82$.

■

Si tenemos más de un estimador insesgado, ¿cuál será el mejor? Dada esta situación es conveniente elegir el que tiene menor varianza.

Ejemplo 5.9

Continuando con el ejemplo anterior, veamos cual de los dos estimadores del parámetro θ es el mejor. Primero calculemos las varianzas de ambos estimadores:

$$\begin{aligned} var(\widehat{\theta}_1) &= var\left(\frac{n+1}{n}\,\text{máx}(X_1, X_2, \ldots, X_n)\right) = \left(\frac{n+1}{n}\right)^2 var\big(\text{máx}(X_1, X_2, \ldots, X_n)\big) \\ &= \frac{\cancel{(n+1)^2}}{n^2}\,\frac{n\theta^2}{\cancel{(n+1)^2}(n+2)} = \frac{\theta^2}{n(n+2)} \qquad (5.3) \\ var(\widehat{\theta}_2) &= var\big(2\overline{X}\big) = 4\,var\big(\overline{X}\big) = 4\,\frac{\theta^2/12}{n} = \frac{\theta^2}{3n} \end{aligned}$$

En (5.3) se utilizó que:

$$var\big(\text{máx}(X_1, X_2, \ldots, X_n)\big) = \frac{n\theta^2}{(n+1)^2(n+2)}$$

la demostración de esta igualdad escapa al alcance de este libro.

Por último, como $n(n+2) \geq 3n$ vale para todo $n \geq 0$, tenemos que: $var(\widehat{\theta}_1) \leq var(\widehat{\theta}_2)$. Por lo tanto, el mejor estimador para el parámetro θ, del Ejemplo 5.8, es $\widehat{\theta}_1$.

■



Observación:

Si alguno de los estimadores no es insesgado existen otros criterios para decidir cuál es el mejor. Pero esto no se tratará en este curso.

Definición:

Se denomina **error estándar** de un estimador a su desviación estándar, es decir, $dt(\widehat{\theta}) = \sqrt{var(\widehat{\theta})}$.

Transformaciones del parámetro

Si $\widehat{\theta}$ es un estimador de θ, y queremos estimar por ejemplo θ^3, es decir, queremos $\widehat{\theta^3}$, parecería razonable usar como estimador $\left(\widehat{\theta}\right)^3$. En general cuando queremos estimar una función *continua* $h(\theta)$, usaremos como estimador $h(\widehat{\theta})$.

Por ejemplo, en la fórmula del error estándar puede haber parámetros desconocidos cuyos valores se pueden estimar, al sustituir dichos parámetros por sus estimadores, se obtiene el **error estándar estimado** del estimador, se suele denotar con $\widehat{dt(\widehat{\theta})}$.

Ejemplo 5.10

Si $X \sim B(n,p)$ ya vimos que $\widehat{p} = X/n$ es un estimador del parámetro p, su error estándar es $dt(\widehat{p}) = \sqrt{var(X/n)} = \sqrt{p(1-p)/n}$. Además, su error estándar estimado es:

$$\sqrt{\frac{\frac{X}{n}\left(1-\frac{X}{n}\right)}{n}} = \sqrt{\frac{\widehat{p}\,(1-\widehat{p})}{n}}$$

■

Ejemplo 5.11

Sea $X_1, X_2, \ldots, X_n$ una m.a. de una distribución con media μ y varianza σ^2. Ya vimos que $\overline{X}$ es un estimador de μ, el error estándar de $\overline{X}$ es $dt(\overline{X}) = \sigma/\sqrt{n}$ y el error estándar estimado de $\overline{X}$ es $S/\sqrt{n}$.

■

EJERCICIO 5.1

1. Se analizaron doce muestras de cierta marca de pan blanco (A) y se determinó el contenido de carbohidratos (expresado en porcentaje), obteniéndose los siguientes valores:

76.93	76.88	77.07	76.68	76.39	75.09
76.88	77.67	78.15	76.50	77.16	76.42

Estime la media, la mediana y el desvío estándar del contenido de carbohidratos para esta marca.

2. Se supone que el tiempo de vida (en horas) de un tipo de lámpara tiene distribución exponencial. Se prueban 10 lámparas de ese tipo y se observa que los tiempos de vida de las mismas son:

7.5	28.2	47.4	17.2	8.5	60.1	21.3	29.5	2.7	5.5

Estime el parámetro λ de la distribución y la probabilidad de que una lámpara de ese tipo dure más de 50 horas.

3. Sea X_1, X_2, ..., X_n una muestra aleatoria de una distribución con media μ y varianza σ^2. Considere los siguientes 3 estimadores para μ:

$$\widehat{\mu}_1 = \frac{X_1 + X_2}{2}, \quad \widehat{\mu}_2 = \frac{X_1}{4} + \frac{X_2 + ... + X_{n-1}}{2(n-2)} + \frac{X_n}{4} \quad \text{y} \quad \widehat{\mu}_3 = \overline{X}$$

   a. Demostrar que los tres estimadores son insesgados para μ.

   b. Calcular sus varianzas y compararlas. ¿Cuál es el mejor?

Intervalos de confianza

En los ejemplos anteriores, hemos estimado un parámetro que puede tomar cualquier valor dentro de un intervalo real, pero sabemos que es prácticamente imposible que nuestra estimación sea exactamente igual al parámetro que deseamos estimar. Por ese motivo, para dar una idea de la precisión de la estimación, se busca dar una estimación mediante un intervalo de confianza.

Definición:

Sea $X_1, X_2, \ldots, X_n$ una m.a. de una distribución $F(\theta)$. Un **intervalo de confianza** de nivel $(1-\alpha)$, o intervalo de $100(1-\alpha)\,\%$ de confianza, es un intervalo de extremos aleatorios, $g_1(X_1, X_2, \ldots, X_n)$ (extremo inferior) y $g_2(X_1, X_2, \ldots, X_n)$ (extremo superior), que contiene al parámetro θ, con probabilidad $1-\alpha$, esto quiere decir que:

$$P\Big(g_1(X_1, X_2, \ldots, X_n) \leq \theta \leq g_2(X_1, X_2, \cdots, X_n)\Big) = 1 - \alpha$$

Notación

Al intervalo de confianza que contiene al parámetro θ, con probabilidad $1-\alpha$, lo denotaremos como $IC_{(1-\alpha)}(\theta)$.



Observación:

Es deseable que el nivel de confianza sea lo mayor posible. Generalmente se utilizan los niveles 0.95 ó 0.99.

¿Cómo construimos un IC?

Los pasos a seguir son:

Paso 1. Se busca una función de la m.a. y del parámetro de interés, que llamaremos **función pivote**, cuya distribución no dependa de ningún parámetro desconocido y que en su expresión el único valor desconocido sea el parámetro de interés. Denotaremos a la función pivote como $h(X_1, X_2, \ldots, X_n, \theta)$.

Paso 2. Determinar un par de números reales a y b, tales que:

$$P\Big(a < h(X_1, X_2, \ldots, X_n, \theta) < b\Big) = 1-\alpha \tag{5.4}$$

Paso 3. Siempre que sea posible, a partir de (5.4), despejar los extremos aleatorios, finalmente obtendremos el intervalo deseado:

$$IC_{(1-\alpha)}(\theta) = \Big(g_1(X_1, X_2, \ldots, X_n), g_2(X_1, X_2, \ldots, X_n)\Big)$$

Veamos como se aplican estos pasos con un ejemplo.

Ejemplo 5.12

Consideremos la distribución de los niveles de colesterol en sangre de los hombres de cierta comunidad, hipertensos y fumadores. Se sabe que esta distribución es aproximadamente normal, se desconoce su media μ, pero se sabe que su desviación típica $\sigma = 46\ mg/100\ ml$, (aunque no se conoce μ se supone que σ es la misma que la de la población de adultos de sexo masculino de esa comunidad). Se desea conocer el nivel medio de colesterol en sangre de este grupo, entonces se seleccionan 12 hombres hipertensos y fumadores y se determina el nivel de colesterol para cada uno. El nivel de colesterol en sangre (medido en $mg/100\ ml$) para cada individuo es una v.a. X_i que tiene distribución normal con media μ (el valor que se desea conocer) y el σ antes mencionado.

Cuando se promedian los 12 valores observados, se obtiene un $\overline{x} = 217\ mg/100\ ml$. Notar que μ es la media "verdadera" desconocida de las observaciones X_i, mientras que $\overline{x}$ es la media de la muestra. Este valor es una estimación de μ.

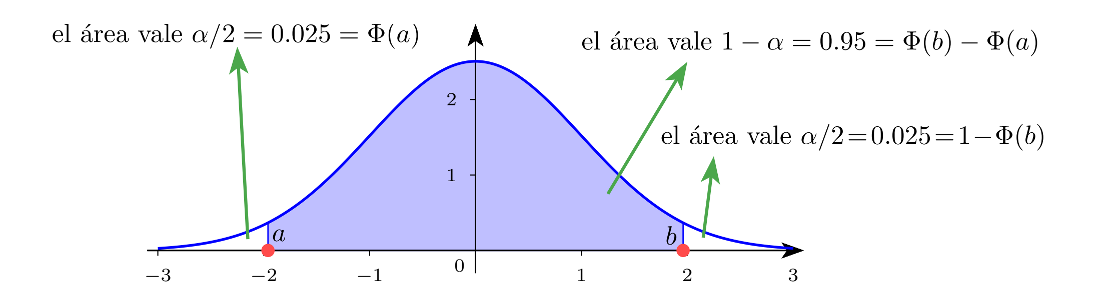
En este caso elegimos $1 - \alpha = 0.95$.

Paso 1. Como las X_i son una m.a. de una $N(\mu, \sigma^2)$, la función

$$Z = h(X_1, X_2, \ldots, X_n, \mu) = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

que tiene distribución $N(0, 1)$, será nuestra función pivote.

Paso 2. Para buscar los valores de a y b que verifiquen (5.4), debemos notar que la función pivote tiene distribución normal estándar, entonces gráficamente tenemos que:



Buscando en la Tabla, vemos que $\Phi(-1.96) = 0.025$ entonces en este ejemplo $a = -1.96$ y $b = 1.96$. Por lo tanto, tenemos que:

$$P\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

Paso 3. Despejando μ en la desigualdad de la probabilidad anterior, obtenemos:

$$P\left(\overline{X} - 1.96 \times \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right) = 0.95,$$

quiere decir que el intervalo obtenido es:

$$IC_{(0.95)}(\mu) = \left(\overline{X} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right) \tag{5.5}$$

Este intervalo de extremos aleatorios contiene al verdadero valor del parámetro μ con probabilidad 0.95, dicho de otro modo, es un intervalo de 95 % de confianza para μ. El intervalo aleatorio, se puede abreviar como $IC_{(0.95)}(\mu) = \overline{X} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$.

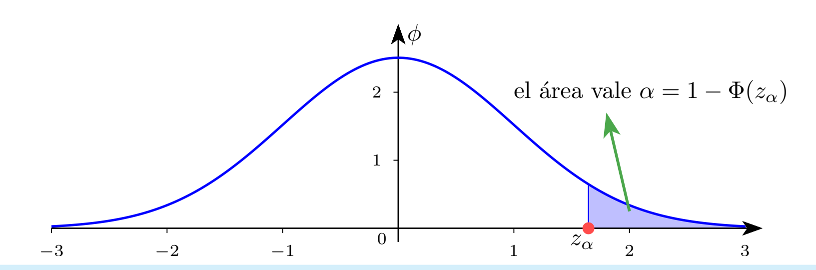
Utilizando los valores del ejemplo y reemplazando $\overline{X}$ por $\overline{x} = 217$, obtenemos:

$$\left(217 - 1.96 \times \frac{46}{\sqrt{12}}, 217 + 1.96 \times \frac{46}{\sqrt{12}}\right) = (191, 243)$$

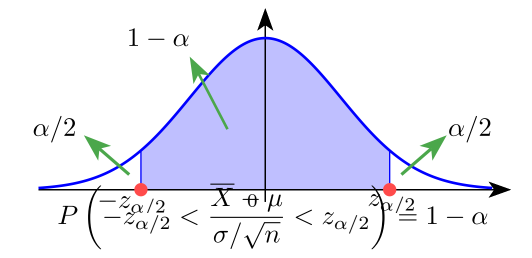
■

Notación

Sea $Z \sim N(0,1)$ y sea α un número real entre 0 y 1, se define el valor crítico, z_α, como el valor tal que $P(Z > z_\alpha) = \alpha$. Informalmente, es el punto sobre el eje x que verifica que el área a su derecha, bajo la curva de densidad, es igual a α. Gráficamente:



El procedimiento que utilizamos para construir un intervalo con un nivel 0.95 para la media de una distribución normal, se puede aplicar para cualquier nivel de confianza $1-\alpha$, en este caso se reemplazan los valores -1.96 y 1.96 por los valores críticos $-z_{\alpha/2}$ y $z_{\alpha/2}$, entonces:



$$P\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$$

y llegamos a:

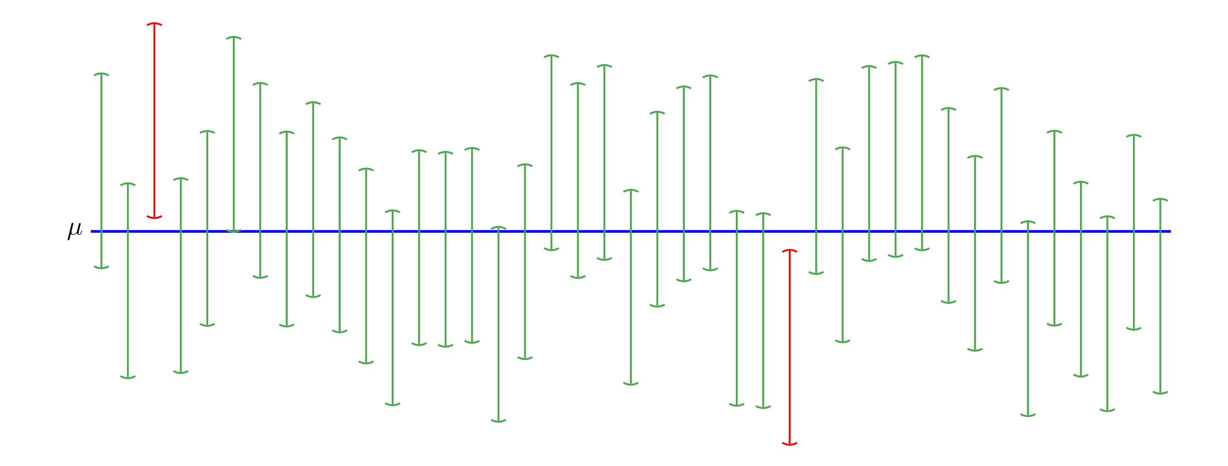
$$P\left(\overline{X} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Finalmente al intervalo:

$$IC_{(1-\alpha)}(\mu) = \left(\overline{X} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\,, \overline{X} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right) \tag{5.6}$$

Interpretación de un intervalo de confianza

En el Ejemplo 5.12, decimos que el nivel de confianza es 0.95, dado que la probabilidad de que el intervalo aleatorio (5.5) contenga al parámetro es 0.95. Es importante recordar que al reemplazar los estadísticos por los valores de la muestra obtuvimos el intervalo de números reales (191, 243), éste ya no es aleatorio y no tiene sentido decir que contiene a μ con probabilidad 0.95. La interpretación correcta del "nivel de confianza" es la siguiente: supongamos, para el ejemplo, que se seleccionan muchas muestras aleatorias de 12 hombres de esa población y se construyen intervalos de confianza utilizando el mismo procedimiento. Con cada muestra de 12 observaciones tendremos un valor de $\overline{x}$ diferente, y en consecuencia un intervalo numérico diferente. Lo que podemos afirmar es que aproximadamente el 95 % de estos intervalos contendrán al verdadero valor μ, y naturalmente habrá aproximadamente un 5 % de dichos intervalos que no contendrán al verdadero valor μ. Entonces cuando construimos un sólo intervalo de nivel 0.95, podemos tener un 95 % de confianza de que ese intervalo sea uno de los que contienen a μ.



En esta gráfica se muestran el verdadero valor de μ (línea horizontal azul) y los $IC_{0.95}(\mu)$ obtenidos con 41 muestras de 12 hombres. Puede verse que algunos pocos (los intervalos en rojo) no contienen a μ.

Nivel de confianza, precisión y tamaño de la muestra en un intervalo de confianza

Como resulta lógico, es deseable que el nivel de confianza $1-\alpha$ sea lo mayor posible, pero el valor de z_α aumenta cuando elegimos valores más grandes para el nivel $1-\alpha$. Por ejemplo si queremos un nivel de 99 % los valores críticos son -2.58 y 2.58. Como consecuencia de ésto aumenta la longitud del intervalo. Esto significa que si se quiere más seguridad hay que pagarla con menos precisión. En el Ejemplo 5.12, si deseamos un nivel de 99 % de confianza, el intervalo será:

$$\left(217 - 2.58 \times \frac{46}{\sqrt{12}}, 217 + 2.58 \times \frac{46}{\sqrt{12}}\right) = (183, 251),$$

la longitud de este intervalo es $L = 251 - 183 = 68$, mientras que al nivel del 95 % la longitud fue de $L = 243 - 191 = 52$.

¿Qué deberíamos hacer si queremos tener un nivel de 99 %, pero mayor precisión, por ejemplo una longitud no mayor de 20? La longitud de (5.6) es $L = 2 \times z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$, entonces reemplazando los valores, $z_{\alpha/2}$, σ y n, planteamos:

$$L = 2 \times 2.58 \times \frac{46}{\sqrt{n}} \leq 20$$

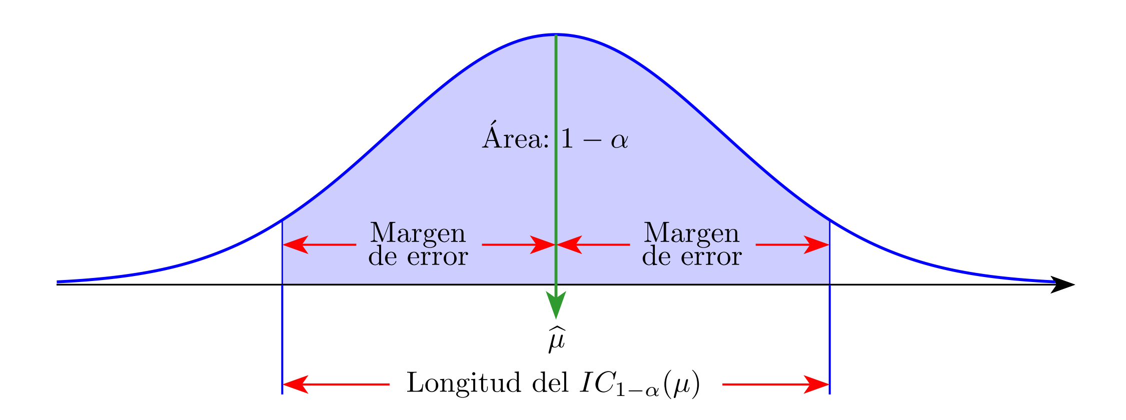
Podemos despejar:

$$\sqrt{n} \geq 2 \times 2.58 \times \frac{46}{20}$$

y obtenemos $n \geq 140.8494$. En consecuencia, necesitaríamos una muestra de por lo menos 141 hombres para lograr un intervalo de 99 % de confianza con longitud no mayor de 20.

En general, a mayor nivel de confianza se tiene menor precisión (un intervalo más largo), y la solución para conseguir el nivel deseado con la precisión deseada es aumentar el tamaño de la muestra.

Para el caso antes analizado, la relación entre longitud del intervalo y error de estimación se ve en la siguiente gráfica:



Intervalos para la media de una población normal con varianza conocida

Para el caso en que la m.a. tenga distribución normal con varianza conocida, como vimos en el Ejemplo 5.12, la función pivote es:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

que tiene distribución N(0,1). Y se obtiene el siguiente intervalo de confianza de nivel $1-\alpha$

$$IC_{1-\alpha}(\mu) = \left(\overline{X} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right)$$

EJERCICIO 5.2

Se repitieron 4 mediciones del valor de creatinina en sangre en una muestra, con un método cuyo desvío estándar es 0.09 mg/dL. Siempre se supone que los errores de medición tienen distribución normal. Se obtuvo una media muestral de 1.03 mg/dL.

1. Calcule un intervalo de confianza del 97 % para el valor de creatinina de la muestra.

2. Calcule la longitud del intervalo de confianza obtenido.

3. ¿Cuántas repeticiones deberían hacerse para que la longitud del intervalo de confianza sea menor que 0.15?

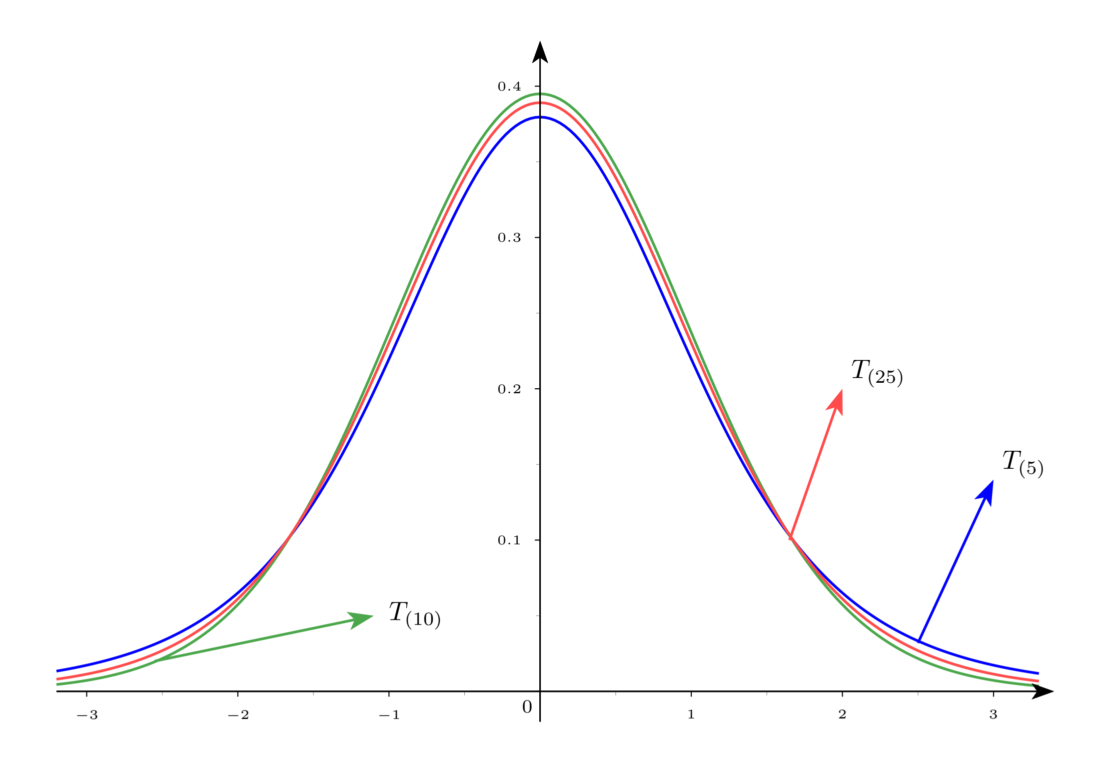
4. Si con esos datos se calculó el intervalo de confianza (0.9418, 1.1182), determine cuál es el nivel de confianza de dicho intervalo.

Intervalos para la media de una población normal con varianza desconocida

En la mayoría de los problemas reales, aún cuando pueda suponerse que la distribución de los datos es aproximadamente normal, la media y la varianza son desconocidas. En ese caso, para construir un intervalo de confianza para la media, no podemos usar la función pivote que usamos anteriormente, porque esa función depende de σ que es desconocido. La función pivote que se debe usar es:

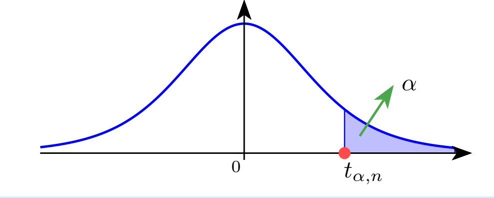
$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

Cuando las X_i son una m.a. de una distribución normal, T tiene distribución de Student con $n-1$ grados de libertad, que se denota como $T_{(n-1)}$. Esta distribución es simétrica, y existen tablas con los valores críticos para ciertos valores de grados de libertad. Las siguientes tres gráficas corresponden a las funciones de densidad para distintos valores de n:



Notación

Sea $T \sim T_{(n)}$ y sea $0 < \alpha < 1$, se define el valor crítico $t_{\alpha,n}$ como el valor que verifica $P(T > t_{\alpha,n}) = \alpha$.



Entonces, siguiendo el procedimiento antes descrito, obtenemos el siguiente intervalo

de $100(1-\alpha)\,\%$ de confianza para parámetro μ

$$IC_{1-\alpha}(\mu) = \left(\overline{X} - t_{\alpha/2,n-1} \times \frac{S}{\sqrt{n}}, \overline{X} + t_{\alpha/2,n-1} \times \frac{S}{\sqrt{n}}\right) \tag{5.7}$$

Ejemplo 5.13

Consideremos las siguientes 7 mediciones de la concentración de ion nitrato (en g/ml) en una muestra de agua:

49 50 51 51 52 53 48

Se desea estimar el valor verdadero μ de la concentración, mediante un intervalo de confianza. Se supone que cada observación X_i es una v.a. con distribución normal con media μ, la que estimamos con la media muestral $\overline{x} = 50.57$.

Como cada X_i es una v.a. con distribución normal, consideramos la función pivote

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

que tiene distribución $T_{(n-1)}$. Reemplazando $\overline{X}$ y S por los valores calculados $\overline{x}$ y s, obtenemos un intervalo real. En nuestro caso, $\overline{x} = 50.57$, $s = 1.72$, y tomemos $1-\alpha = 0.95$ (el nivel de confianza es $95\,\%$), se busca en la tabla el valor correspondiente a grados de libertad $n-1=6$ y $\alpha/2 = 0.025$, que es $t_{0.025,6} = 2.4469$.

El intervalo es

$$\left(50.57 - 2.4469 \times \frac{1.72}{\sqrt{7}}, 50.57 + 2.4469 \times \frac{1.72}{\sqrt{7}}\right) = (48.98, 52.16)$$

■

EJERCICIO 5.3

Se midieron las tallas (en cm) a los 12 meses de edad de 16 niñas con hipotiroidismo congénito (HC). Se obtuvieron los siguientes valores $\overline{x} = 73.85$ y $s = 2.58$. Se puede suponer que la talla es una variable aleatoria con distribución normal.

1. Construya un intervalo de $98\,\%$ de confianza para la talla media a los 12 meses de edad de las niñas con HC.

2. Calcule la longitud del intervalo de confianza obtenido.

Intervalos de confianza para la media con muestras grandes

Para construir el intervalo de confianza (5.7) nos basamos en la suposición de que la distribución de la población era normal. Si ése no es el caso, la función pivote utilizada no tendría distribución de Student. Cuando no conocemos la distribución de los datos, es necesario usar algún tipo de

aproximación. Recordemos el Teorema Central del Límite (TCL) que vimos en el Capítulo 4, según este teorema, si tenemos una m.a. $X_1, X_2, \ldots, X_n$ de cualquier distribución y n es suficientemente grande, la distribución de

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

se aproxima a una $N(0,1)$. Se puede demostrar que si se reemplaza σ por S, la distribución también se aproxima a una $N(0,1)$. Este resultado es el que usaremos cuando no conocemos la distribución de los datos. El procedimiento es el mismo, partimos de la función pivote

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

que, considerando que n es grande, tiene una distribución aproximadamente $N(0,1)$. Entonces los valores críticos que elegimos son $-z_{\alpha/2}$ y $z_{\alpha/2}$ y podemos afirmar que:

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{S/\sqrt{n}} < z_{\alpha/2}\right) \cong 1 - \alpha$$

Ejemplo 5.14

La contaminación de metales pesados de varios ecosistemas es una amenaza ambiental. Un artículo científico reporta que, para una muestra de $n = 56$ peces de la especie Mugil liza, la concentración media muestral de zinc en el hígado fue de $9.15\,\mu g/g$ y la desviación estándar muestral fue de $1.27\,\mu g/g$. Se desea estimar la concentración media poblacional de zinc en el hígado de esa especie de peces, mediante un intervalo de 95 % de confianza.

Como n es suficientemente grande, la distribución de

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

se aproxima a una $N(0,1)$. Reemplazando con los datos del ejemplo, $\overline{x} = 9.15$, $s = 1.27$, y $z_{0.025} = 1.96$, obtenemos: $(8.82, 9.48)$, este intervalo tiene nivel de confianza aproximado de 95 %.

Ahora, si deseamos hallar el valor de n para el cual la longitud del intervalo del 95 % de confianza sea a lo sumo 0.5, el planteo sería:

$$2 \times z_{0.025} \times \frac{S}{\sqrt{n}} \leq 0.5 \tag{5.8}$$

Observemos que S también depende de n. En estos casos, se reemplaza el valor de S de alguna muestra previa que se tenga. En nuestro caso, en (5.8) nos quedará que

$$2 \times 1.96 \times \frac{1.27}{\sqrt{n}} \leq 0.5 \quad \Rightarrow \quad n \geq \left(\frac{2 \times 1.96 \times 1.27}{0.5}\right)^2 = 99.1379$$

Entonces, el valor de n debe ser al menos 100.

■

EJERCICIO 5.4

En un estudio nutricional se evaluó el consumo diario de calorías en un grupo de 40 adolescentes de sexo femenino. La media y desviación típica muestrales de esos valores, en kilocalorías por kilogramos, fueron $\overline{x} = 32.85$ y $s = 5.76$. No hay evidencias de que el consumo diario de calorías siga una distribución normal.

1. Construya un intervalo de aproximadamente 95 % de confianza para la media del consumo diario de calorías para la población de adolescentes.

2. Si se desea que la longitud del intervalo de confianza no sea mayor que 3, ¿cuántas adolescentes es necesario encuestar?

Intervalos de confianza con muestras grandes para una proporción

Sea X una v.a. con distribución binomial con parámetros n y p, por lo tanto:

$$E(X) = np \quad \text{y} \quad dt(X) = \sqrt{np(1-p)}$$

Ya vimos que $\widehat{p} = \dfrac{X}{n}$, la proporción observada en la muestra, es un estimador insesgado de p y cumple:

$$E(\widehat{p}) = p \quad \text{y} \quad dt(\widehat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

Con el caso particular del TCL para la binomial, sabemos que la distribución de

$$\frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \quad \text{se aproxima a una } N(0,1)$$

También vale que la distribución de

$$\frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \quad \text{se aproxima a una } N(0,1)$$

Entonces eligiendo los valores críticos $-z_{\alpha/2}$ y $z_{\alpha/2}$, se cumple:

$$P\left(-z_{\alpha/2} < \frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} < z_{\alpha/2}\right) \cong 1 - \alpha$$

Luego, se puede obtener un intervalo de confianza para p con nivel aproximadamente $1 - \alpha$ (para n grande), de la forma:

$$IC_{1-\alpha}(p) = \left(\widehat{p} - z_{\alpha/2} \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} + z_{\alpha/2} \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right)$$

y abreviado es:

$$IC_{1-\alpha}(p) = \widehat{p} \pm z_{\alpha/2} \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

Aclaración

En la práctica, el extremo inferior del intervalo podría dar negativo, en cuyo caso se lo hace igual a cero; de igual forma, si el extremo superior da mayor que 1, se lo hace igual a 1.

Ejemplo 5.15

Se realizó un estudio para detectar anemia en niños menores de 6 años en una comunidad rural. Se seleccionaron al azar 230 niños de esa comunidad y se encontraron 107 con anemia ($Hg < 11\, g/dl$). Se desea estimar mediante un intervalo de confianza el porcentaje de niños con anemia en esa comunidad. El número de casos, en la muestra de 230, con anemia es $x = 107$.

Definimos a la variable aleatoria X como la cantidad de niños menores de 6 años con anemia entre 230. Luego X es una v.a. con distribución binomial con parámetros 230 y p, luego consideramos la función pivote

$$\frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \quad \text{que se aproxima a una } N(0,1)$$

En nuestro caso, $\widehat{p} = 0.4652$, y si elegimos $1 - \alpha = 0.95$, es $z_{\alpha/2} = 1.96$, luego el intervalo resulta

$$(0.4007, 0.5297) \tag{5.9}$$

■

Conociendo el tamaño de la población, se puede construir un intervalo de confianza para la cantidad de individuos en esa población que tienen la característica que se está estudiando. En el Ejemplo 5.15, si se desea evaluar los costos de un programa de intervención para mejorar la salud comunitaria, interesa conocer el número de niños con anemia.

Llamemos N a la cantidad de niños menores de 6 años en la población, y M a la cantidad de niños con anemia, la verdadera proporción de niños con anemia se define como $p = M/N$, y de allí $M = Np$, entonces se puede estimar $\widehat{M} = \widehat{p}N$. En el ejemplo anterior, si $N = 1500$, $\widehat{M} = 0.4652 \times 1500 = 697.8$.

Para construir un intervalo de confianza para M se puede proceder como sigue: sabemos que

$$P\left(\widehat{p} - z_{\alpha/2} \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} < p < \widehat{p} + z_{\alpha/2} \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) \cong 1 - \alpha$$

entonces si multiplicamos por N:

$$P\left[N\left(\widehat{p} - z_{\alpha/2} \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) < Np < N\left(\widehat{p} + z_{\alpha/2} \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right)\right] \cong 1 - \alpha$$

obtendremos el intervalo de nivel $1 - \alpha$ para M que es:

$$IC_{1-\alpha}(M) = \left(N\left(\widehat{p} - z_{\alpha/2} \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right), N\left(\widehat{p} + z_{\alpha/2} \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right)\right)$$

En el ejemplo, si a los extremos del intervalo (5.9) los multiplicamos por el valor de N, obtendremos $(601.05, 794.55)$, pero recordando que M es un número natural, finalmente obtenemos $(601, 795)$.

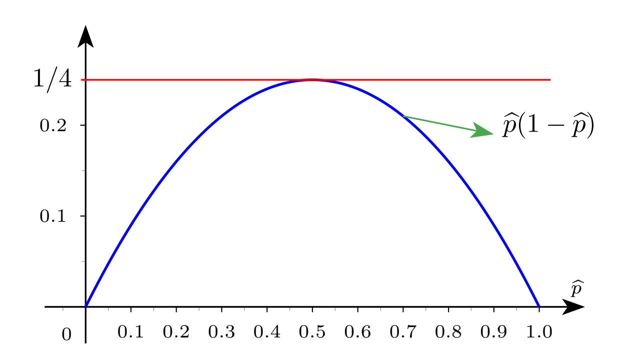
Nivel de confianza, precisión y tamaño de la muestra

En el Ejemplo 5.15, la longitud del intervalo para la proporción de niños con anemia, es 0.129. En general, la longitud es:

$$L = 2 \times z_{\alpha/2} \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \tag{5.10}$$

Si se pretende estimar la proporción de niños anémicos con un error de estimación no mayor del 5 %, esto quiere decir que la longitud del intervalo no debe ser mayor que 0.10, antes de realizar el estudio se debería determinar cuántos niños o cuántas muestras de sangre se necesitarán analizar.

El problema en este caso, es que la longitud del intervalo (5.10) depende también de $\widehat{p}$ que no se conoce antes del estudio. Pero se puede ver fácilmente que para cualquier $\widehat{p}$ vale, $\widehat{p}(1-\widehat{p}) \leq 1/4$ como se muestra en la siguiente gráfica:



Entonces:

$$L = 2 \times z_{\alpha/2} \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq 2 \times z_{\alpha/2} \times \sqrt{\frac{1/4}{n}} = \frac{z_{\alpha/2}}{\sqrt{n}}$$

Luego si queremos que $L \leq d$, debemos hacer $z_{\alpha/2}/\sqrt{n} \leq d$ y de allí podemos despejar el valor de n necesario para que la longitud del intervalo sea a lo sumo d.

Para el ejemplo anterior:

$$L = 2 \times 1.96 \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \leq \frac{1.96}{\sqrt{n}} \leq 0.10$$

luego $n \geq (1.96/0.10)^2 = 384.16$ entonces si n es al menos 385, nos aseguramos que la longitud del intervalo será menor de 0.10.

EJERCICIO 5.5

Una de las metas de un programa de pesquisa neonatal de hipotiroidismo congénito, es lograr la detección de la enfermedad en los primeros días de vida, por ese motivo el protocolo de la pesquisa indica que la muestra de sangre para el análisis debe ser tomada en los primeros 5 días de vida. Se quiere construir un intervalo de confianza de nivel aproximado 0.95, para la proporción de casos donde no se cumple el protocolo, analizando los registros del programa.

1. Si se desea que la longitud del intervalo de confianza para esa proporción sea menor que 0.05, ¿cuántos registros se deberían observar?

2. Si se supone que dicha proporción es menor que 0.20, y se desea que la longitud del intervalo de confianza sea menor que 0.05, ¿cuántos registros se deberían observar?

3. Se eligieron al azar 300 registros de ese programa y se observó que en 54 casos la muestra había sido tomada después de los 5 días de vida. Construya el intervalo de confianza para la proporción de casos en que no se cumple el protocolo.

4. Si este programa se aplica a todos los recién nacidos en una región, donde hay aproximadamente 10000 nacimientos por año. Construya un intervalo de confianza para el número de niños a los que se les realiza la toma de muestra después del tiempo especificado.

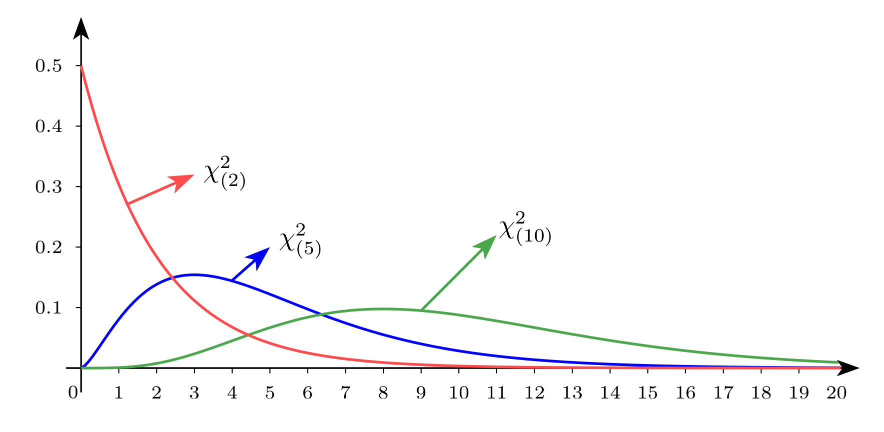
Intervalos para la varianza de una distribución normal

Hay situaciones en que interesa hacer inferencias sobre la varianza o la desviación típica, por ejemplo cuando queremos conocer la precisión de un método de medición.

Si tenemos una m.a. de una distribución normal y queremos calcular un intervalo de confianza para la varianza, la función pivote $h(X_1, X_2, \dots, X_n, \sigma^2)$, que usaremos es:

$$V = \frac{(n-1)S^2}{\sigma^2}$$

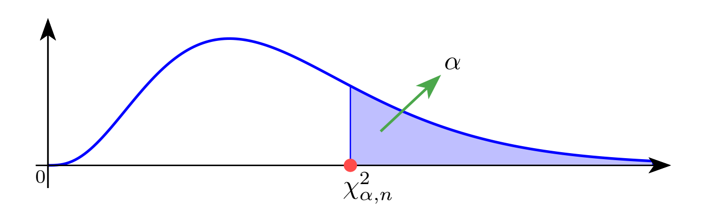
ya que puede demostrarse que, cuando las X_i tienen distribución $N(\mu, \sigma^2)$, V tiene distribución Chi-cuadrado con $n-1$ grados de libertad, $\chi^2_{(n-1)}$. Esta distribución no es simétrica, la densidad es no nula sólo para $x > 0$.



También existen tablas para los valores críticos de esta distribución, para diferentes valores de grados de libertad.

Notación

Sea $U \sim \chi^2_{(n)}$ y sea $0 < \alpha < 1$, se define el valor crítico $\chi^2_{\alpha,n}$ como el valor que verifica $P(U > \chi^2_{\alpha,n}) = \alpha$.



Como siempre, necesitamos un par de valores tales que V se encuentre entre ellos con probabilidad $1-\alpha$. Pero esta distribución no es simétrica, entonces deberemos elegir los valores $\chi^2_{1-\alpha/2,n-1}$ y $\chi^2_{\alpha/2,n-1}$ tales que:

$$P\left(\chi^2_{1-\alpha/2,n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right) = 1-\alpha$$

y finalmente el intervalo:

$$IC_{1-\alpha}(\sigma^2) = \left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right)$$

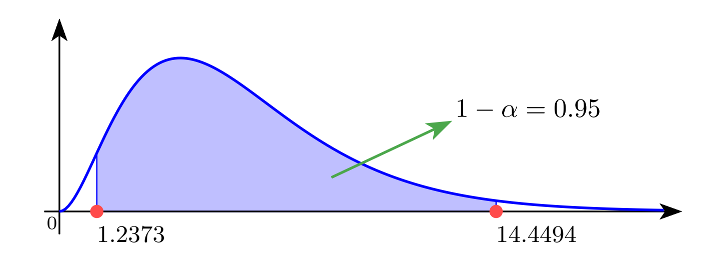
de extremos aleatorios. Como siempre, esto significa que el verdadero valor de σ^2 se encuentra en ese intervalo con probabilidad $1-\alpha$. Reemplazando el estimador S^2 por el valor de la muestra s^2, obtenemos un intervalo numérico.

Ejemplo 5.16

Volviendo a los datos del Ejemplo 5.13, consideramos la función pivote:

$$V = \frac{(n-1)S^2}{\sigma^2}$$

Cuando las X_i tienen distribución $N(\mu, \sigma^2)$, V tiene distribución Chi-cuadrado con $n-1$ grados de libertad. Luego tenemos que $s = 1.718$ y eligiendo $1-\alpha = 0.95$, los valores críticos los buscamos en la tabla de la Chi-cuadrado con $n-1 = 6$ grados de libertad, obteniendo $\chi^2_{0.025,6} = 14.4494$ y $\chi^2_{0.975,6} = 1.2373$



Finalmente, el intervalo para σ^2 es:

$$\left(\frac{6 \times 1.718^2}{14.4494}, \frac{6 \times 1.718^2}{1.2373}\right) = (1.23, 14.31)$$

Si deseamos un intervalo para σ, debemos sacar raíz cuadrada a cada extremo del intervalo anterior, así obtenemos $(1.11, 3.78)$.

■

EJERCICIO 5.6

Usando los datos del Ejercicio 5.3, calcular intervalos de confianza de nivel 0.95 para la varianza y el desvío estándar de la distribución de tallas en la población estudiada.

Referencias

Agresti, A. & Franklin, C. A. (2009). *Statistics: The Art and Science of learning from Data*. Pearson New International edition.

Altman, D. G. (1990). *Practical Statistics for Medical Research*. Published by Chapman & Hall.

Daniel, W. (2002). *Bioestadística: Base para el análisis de las ciencias de la salud*. Ed. Limusa Wiley.

Devore Jay, L. (2001). *Probabilidad y Estadística para Ingeniería y Ciencias*. Ed. Books/Cole Publishing Company.

Dixon, W. & Massey, F. (1970). *Introducción al Análisis Estadístico*. México. Libros Mc Graw-Hill.

Maronna, R. (1995). *Probabilidad y Estadística Elementales para Estudiantes de Ciencias*. Buenos Aires.Ed. Exactas.

Mendenhall, W., Beaver, R. J. & Beaver, B. M. (2006). *Introducción a la Probabilidad y Estadística*. México. Cengage Learning Editores.

Ross, S. M. (1987). *Introduction to Probability and Statistics for Enginees and Scientists*. Published by John Wiley & Sons.

Wackerly, D. D., Mendenhall, W. & Scheaffer, R. L. (2010). *Estadística Matemática con aplicaciones*. México. Cengage Learning Editores.

Walpole, R. E. & Myers, R. H. (2007). *Probabilidad y Estadística para Ingeniería y Ciencias*. México. Ediciones McGraw-Hill.

CAPÍTULO 6

Tests de hipótesis

Introducción

En el capítulo anterior vimos cómo estimar un parámetro con un solo número (una estimación puntual) o un intervalo completo de valores plausibles (un intervalo de confianza), a partir de los datos muestrales. Con frecuencia, sin embargo, el objetivo de una investigación no es estimar un parámetro sino decidir cuál de dos pretensiones contradictorias sobre él es la correcta. Los métodos para lograr esto constituyen la parte de la inferencia estadística llamada pruebas o tests de hipótesis.

En muchos aspectos, el procedimiento formal para pruebas de hipótesis es semejante al método científico. Observar la naturaleza, formular una teoría y confrontarla con lo observado.

En nuestro contexto, tenemos una hipótesis sobre la población (por ejemplo acerca de su media o de su varianza) y queremos saber si es cierta o no, para verificarla tomamos una muestra aleatoria, y en función de la misma, decidimos si la aceptamos o no.

Ahora bien, ¿cómo se utiliza la estadística en este procedimiento? Probar una hipótesis requiere tomar una decisión cuando se compara la muestra observada contra la teoría propuesta. ¿Cómo decidimos si la muestra concuerda o no con la hipótesis del científico? ¿Cuándo debemos rechazar la hipótesis y cuándo debemos aceptarla? ¿Cuál es la probabilidad de que tomemos una mala decisión? Y, en particular, ¿qué función de las mediciones muestrales debe emplearse para llegar a una decisión? En este capítulo trataremos de dar respuesta a estas preguntas.

Test de hipótesis para la media de una distribución normal con varianza conocida

Para presentar las ideas de test de hipótesis, comencemos desarrollando un ejemplo:

Una concentración de mercurio (Hg) en el agua igual a 1 mcg/l, se considera un riesgo para la salud. Se realizan 6 determinaciones de concentración de Hg en una muestra de agua y se obtienen los siguientes valores (mcg/l):

$$0.86 \quad 0.89 \quad 0.93 \quad 1.02 \quad 0.96 \quad 0.83$$

Suponemos que el método tiene un error de medición con $\sigma = 0.08\ mcg/l$. Queremos determinar si esa muestra de agua puede considerarse libre de contaminación. Para ello, debemos decidir entre estas dos hipótesis:

- La concentración de Hg en la muestra de agua es admisible.
- La concentración de Hg en la muestra de agua es un riesgo para la salud.

Podemos modelizar esta situación como sigue: tenemos una m.a. $X_1, X_2, \ldots, X_6$ donde cada X_i es el resultado de la i-ésima determinación y tiene distribución $N(\mu, 0.08^2)$. El problema a resolver es decidir cuál de las siguientes hipótesis es cierta:

$$H_0 : \mu = 1 \quad \text{(el agua es un riesgo para la salud)}$$
$$H_A : \mu < 1 \quad \text{(el agua es admisible)}$$

donde H_0 se llama **hipótesis nula** y H_A **hipótesis alternativa**.

Debemos notar que al decidirnos por una de las dos hipótesis, pueden ocurrir alguna de estas cuatro situaciones:

Decisión a partir del test	**H_0 es cierta**	**H_0 no es cierta**
Se rechaza H_0	*error tipo I*	correcto
No se rechaza H_0	correcto	*error tipo II*

Como se aprecia en la tabla anterior, pueden cometerse dos tipos de errores, que se los distingue con los nombres de **error de tipo I** y **error de tipo II**.

En nuestro ejemplo el error de tipo I sería afirmar que la muestra de agua tiene una concentración de mercurio aceptable ($\mu < 1$), cuando en realidad la concentración es alta. Y el error de tipo II sería afirmar que la muestra de agua tiene una concentración alta ($\mu = 1$), cuando en realidad es más baja. En este caso, el error de tipo I es más grave que el error de tipo II.

Los procedimientos que vamos a ver, nos permiten acotar la probabilidad de cometer un error de tipo I.

Recordemos que nunca conocemos cuánto vale μ y sólo podemos hacer inferencias, acerca de su valor, basadas en la muestra. Sabemos que $\overline{X}$ es un estimador de μ y el valor $\overline{x}$ observado es una estimación del verdadero μ. Entonces, si $\overline{x}$ resulta mucho más chico que 1, tendremos motivos para pensar que en realidad $\mu < 1$ (cuanto menor sea $\overline{x}$, mayor será la evidencia contra H_0 a

favor de H_A). Debemos decidir cuándo consideramos que $\overline{x}$ es lo suficientemente pequeño como para rechazar la hipótesis nula, manteniendo acotada la probabilidad de error de tipo I. Para esto debemos considerar un estadístico (estadístico de prueba) con distribución conocida cuando H_0 es verdadera y definir una zona de rechazo. En nuestro ejemplo, usaremos el estadístico de prueba:

$$Z = \frac{\overline{X} - 1}{0.08/\sqrt{6}} = \frac{\sqrt{6}\,(\overline{X} - 1)}{0.08}$$

que bajo el modelo propuesto y cuando H_0 es verdadera, tiene distribución $N(0, 1)$.

Supongamos que queremos que la probabilidad de error de tipo I sea 0.05. Se puede establecer una regla de decisión como la siguiente: rechazar H_0 cuando el valor del estadístico de prueba sea menor que -1.645; de este modo nos aseguramos que:

$$P(\text{ error de tipo I }) = P\left(\frac{\sqrt{6}\,(\overline{X} - 1)}{0.08} < -1.645 \middle| H_0 \text{ es verdadera}\right) = \Phi(-1.645) = 0.05$$

En general, para que la probabilidad de error de tipo I sea α la regla será:

$$\text{rechazar } H_0 : \mu = 1 \text{ cuando } z = \frac{\sqrt{6}\,(\overline{x} - 1)}{0.08} < -z_\alpha$$

ya que:

$$P(\text{ error de tipo I }) = P\left(\frac{\sqrt{6}\,(\overline{X} - 1)}{0.08} < -z_\alpha \middle| H_0 \text{ es verdadera}\right) = \Phi(-z_\alpha) = \alpha$$

Este valor α se llama **nivel de significación**. Al fijar un nivel de significación $\alpha = 0.05$, nos aseguramos que la probabilidad de cometer un error de tipo I, no puede ser mayor que 0.05. Para este test, llamamos zona de rechazo a los valores del estadístico de prueba menores que $-z_\alpha$, puede verse que el área bajo la curva de densidad en la zona de rechazo es igual a α.

Con los datos del ejemplo, al reemplazar $\overline{X}$ por $\overline{x} = 0.915$, el valor que toma el estadístico es: $\frac{\sqrt{6}\,(0.915-1)}{0.08} = -2.60$, este valor cae en la zona de rechazo, de modo que podemos rechazar H_0 con nivel $\alpha = 0.05$. Es decir, podemos afirmar que la concentración de Hg en la muestra de agua está dentro de los niveles admisibles, y la probabilidad de equivocarnos al hacer esta afirmación es a lo sumo 0.05. También se dice que el resultado es significativo al 5 %.

También podemos razonar de esta manera: si $\mu = 1$, ¿cuál es la probabilidad de obtener una media muestral que fuera tanto o más pequeña que la observada?, o lo que es equivalente, ¿cuál es la probabilidad de que el estadístico de prueba sea menor o igual que -2.60, si $\mu = 1$? Esta probabilidad puede calcularse ya que el estadístico de prueba tiene distribución $N(0, 1)$ cuando $\mu = 1$ y es:

$$P(Z < -2.60) = \Phi(-2.60) = 0.0047$$

Esto es lo que se llama el "valor-p", cuanto menor sea este p, más evidencia tengo contra H_0. En nuestro ejemplo, valor-$p = 0.0047$ es una probabilidad muy pequeña, podemos rechazar H_0 y afirmar la hipótesis alternativa, es decir que $\mu < 1$ con fuerte convicción.

El valor-p se puede definir como el menor nivel de significación (el más exigente) para el cual se puede rechazar H_0 con los valores observados. Otra manera de expresar la regla de decisión es: se rechaza H_0 cuando el valor-p es menor que el α elegido.

En este caso el valor-$p = 0.0047$, esto significa que podemos rechazar H_0 con cualquier $\alpha \geq 0.0047$ que es equivalente a decir que el resultado es significativo al $0.47\,\%$.

Cualquier prueba de hipótesis estadística funciona de la misma forma y está compuesta de los mismos elementos esenciales.

Un **test o prueba de hipótesis** es un proceso de decisión que, en función de los datos de una m.a. $X_1, X_2, \ldots, X_n$, nos permite decidir entre la validez de dos hipótesis contradictorias.

Los elementos de un test de hipótesis son:

1. Las hipótesis: hipótesis nula e hipótesis alternativa.

2. El estadístico de prueba.

3. La región de rechazo.

Definiremos a continuación formalmente estos elementos:

Definición:

Las hipótesis **nula** y **alternativa** son hipótesis contradictorias sobre el parámetro de interés. Se decide cuál es la hipótesis alternativa, de modo que el error de tipo I sea el más grave (ya que la probabilidad de cometer dicho error quedará acotada). Habitualmente, en los trabajos de investigación, la hipótesis alternativa es la que el investigador desea demostrar.

Definición:

El **estadístico de prueba** es una función de la muestra aleatoria, y debe tener una distribución conocida cuando el valor del parámetro sea el que indica la hipótesis nula. La decisión estadística estará basada en el valor del estadístico de prueba.

Definición:

La **región de rechazo** especifica los valores del estadístico de prueba para los cuales la hipótesis nula ha de ser rechazada a favor de la hipótesis alternativa. Se define de modo que la probabilidad de cometer error de tipo I sea igual a α. Este valor α se llama **nivel de significación** del test, y es elegido "a priori" por el investigador (generalmente 0.05). El área de la zona de rechazo es igual al nivel de significación (α) elegido.

Entonces con todos estos elementos, para una muestra particular, si el valor calculado del estadístico de prueba cae en la región de rechazo, rechazamos la hipótesis nula H_0 y aceptamos la hipótesis alternativa H_A, sabiendo que la probabilidad de equivocarnos (error de tipo I) es α. Si el valor del estadístico de prueba no cae en la región de rechazo, concluimos que a partir de los datos, no tenemos evidencia suficiente para rechazar H_0; si nos equivocamos estaríamos cometiendo un error de tipo II, la probabilidad de error de tipo II se suele llamar β y es más difícil de calcular.

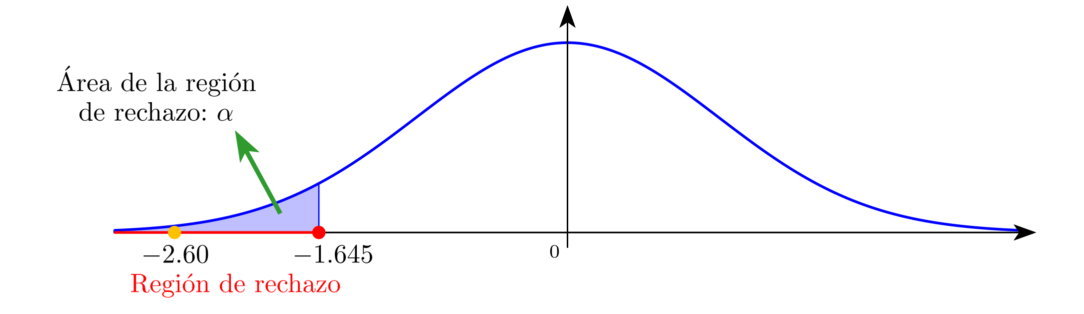
Ejemplo 6.1

Para el ejemplo de la introducción, los elementos del test de hipótesis son:

1. Las hipótesis: $H_0 : \mu = 1$ vs. $H_A : \mu < 1$. Este ejemplo corresponde a un test unilateral pues la alternativa sólo puede ocurrir en una dirección (hacia la izquierda).

2. El estadístico de prueba: $Z = \dfrac{\sqrt{6}\,(\overline{X} - 1)}{0.08}$ y cuando H_0 es verdadera, $Z \sim N(0, 1)$.

3. La región de rechazo: $z = \dfrac{\sqrt{6}\,(\overline{x} - 1)}{0.08} < -z_{0.05} = -1.645$, gráficamente:



En el gráfico se puede ver que el estadístico de prueba $z = -2.60$ cae dentro de la región de rechazo.

■

Definición:

Se define el **valor-*p*** como la probabilidad de que el estadístico de prueba tome un valor "tan extremo" como el obtenido, si fuera cierta la hipótesis nula. O también puede definirse como el menor nivel de significación α, para el cual se puede rechazar la hipótesis nula con los valores observados.



Observación:

Las dos definiciones de anteriores de valor-p son equivalentes. Es importante observar que, el valor-p y el nivel de significación son cosas diferentes: el valor-p depende de los valores observados, mientras que el nivel de significación se elige a priori.

Si el valor-p es menor o igual a un nivel de significancia α asignado previamente, entonces la hipótesis nula puede ser rechazada con dicho α y se puede afirmar que los resultados son estadísticamente significativos a ese nivel.

El Ejemplo 6.1 es un test unilateral, analicemos ahora un ejemplo donde la hipótesis alternativa es bilateral.

Ejemplo 6.2

Una importante propiedad de un método analítico es que no tenga error sistemático. Ya vimos que el resultado de una medición puede modelizarse como una variable aleatoria con distribución normal, cuya media es igual al verdadero valor del analito, cuando el método no tiene error sistemático. Un error sistemático implicaría que esa media difiere del valor verdadero del analito. Para ver si se verifica esta propiedad, se analiza con dicho método una muestra estándar con valor conocido del analito. En este ejemplo, se quiere determinar si un método para medir selenourea en agua tiene error sistemático. Para ello, se analizaron muestras de agua de grifo adicionadas con 50 ng/ml de selenourea. Se sabe que la desviación típica del error de medición es 0.9 ng/ml. Se obtuvieron los siguientes valores (medidos en ng/ml):

48.8 50.2 50.1 48.1 48.7 49.4

En este ejemplo queremos decidir entre estas dos hipótesis:

- el método de medición no tiene error sistemático
- el método de medición tiene error sistemático

Podemos plantear el siguiente modelo probabilístico:

Sea $X_1, X_2, \ldots, X_n$ una m.a. donde X_i es el resultado de la i-ésima medición del contenido de selenourea en la muestra de agua y la suponemos con distribución $N(\mu, 0.9^2)$. En este caso $n = 6$. Si el método no tiene error sistemático, la media μ es igual al verdadero contenido de selenourea en esa muestra. Con este modelo las hipótesis se escriben:

$$H_0 : \mu = 50 \quad \text{(el método de medición no tiene error sistemático)}$$
$$H_A : \mu \neq 50 \quad \text{(el método de medición tiene error sistemático)}$$

Usaremos el estadístico de prueba:

$$Z = \frac{\sqrt{6}\,(\overline{X} - 50)}{0.9}$$

que, bajo el modelo propuesto y cuando H_0 es verdadera, tiene distribución $N(0, 1)$.

En este caso, parece lógico rechazar H_0 cuando $\overline{x}$ sea mucho mayor o mucho menor que 50. La regla de decisión será: rechazar H_0 cuando el valor absoluto del valor del estadístico de prueba sea mayor que 1.96, de este modo nos aseguramos que:

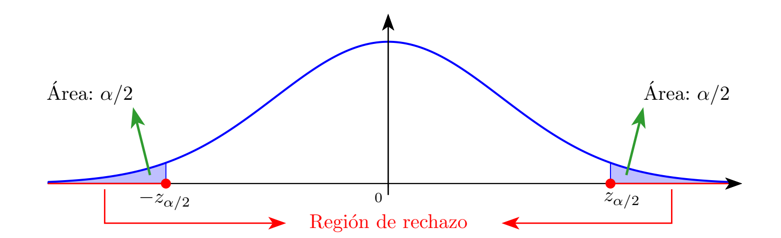
$$P(\text{ error de tipo I }) = P\left(\frac{\sqrt{6}\,|\overline{X} - 50|}{0.9} > 1.96 \middle| H_0 \text{ es verdadera}\right) = 2\big(1 - \Phi(1.96)\big) = 0.05$$

En general, en un test bilateral, para que la probabilidad de error de tipo I sea α la regla será:

$$\text{rechazar } H_0 : \mu = \mu_0 \text{ cuando } \frac{\sqrt{n}\,|\overline{x} - \mu_0|}{\sigma} > z_{\alpha/2}$$

entonces:

$$P(\text{ error de tipo I }) = P\left(\frac{\sqrt{n}\,|\overline{X} - \mu_0|}{\sigma} > z_{\alpha/2} \middle| H_0 \text{ es verdadera}\right) = 2\big(1 - \Phi(z_{\alpha/2})\big) = \alpha$$



En nuestro ejemplo, con las 6 mediciones se obtuvo $\overline{x} = 49.217$, reemplazando $\overline{X}$ por $\overline{x} = 49.217$, el estadístico de prueba toma el valor $\frac{\sqrt{6}\,(49.217-50)}{0.9} = -2.13$. Como este valor cae en la región de rechazo, pues $|-2.13| > 1.96$, podemos rechazar la hipótesis nula con nivel 0.05. Esto significa que podemos afirmar que $\mu \neq 50$, es decir, el método tiene error sistemático y la probabilidad de equivocarnos al hacer esta afirmación es a lo sumo 0.05.

También podemos calcular el valor-p:

$$\text{valor-}p = P\left(\frac{\sqrt{6}\,|\overline{X}-50|}{0.9} > |-2.13| \middle| H_0 \text{ es verdadera}\right) = P\big(|Z| > 2.13\big)$$

$$= P\big(Z > 2.13\big) + P\big(Z < -2.13\big) = 2\big(1 - \Phi(2.13)\big) = 0.0332$$



Observación:

Los test bilaterales son más conservadores que los unilaterales, ya que para un mismo valor del estadístico de prueba, el valor-p es mayor para un test bilateral que para un test unilateral.

Podemos resumir lo que hemos visto sobre tests para la media μ de la siguiente manera:

RESUMEN 6.1

Para la m.a. $X_1, X_2, \ldots, X_n$ con distribución normal con σ_0 conocido.			
Hipótesis nula	$H_0 : \mu = \mu_0$		
Valor del estadístico de prueba	$z = \frac{\sqrt{n}\,(\overline{x} - \mu_0)}{\sigma_0}$		
Hipótesis alternativa	$H_A : \mu < \mu_0$	$H_A : \mu > \mu_0$	$H_A : \mu \neq \mu_0$
Región de rechazo con nivel α	$z < -z_\alpha$	$z > z_\alpha$	$\|z\| > z_{\alpha/2}$

EJERCICIO 6.1

Se extrajo una m.a. de 16 informes de urgencias de los archivos de un servicio de ambulancias. El tiempo medio para que las ambulancias llegaran a sus destinos fue de 13 min. Suponga que la población de tiempos sigue una distribución normal con varianza 9. ¿Es posible concluir, con un nivel de significancia de 0.05, que la media de la población es menor que 10 min?

Test de hipótesis para la media de una distribución normal con varianza desconocida

Cuando la distribución de los datos es normal, pero desconocemos el valor de σ, no podemos usar el mismo estadístico de prueba que en el caso anterior. Recordemos lo que vimos al construir intervalos de confianza, en ese caso usamos una función pivote con distribución de Student. En el caso de un test, cuando $\mu = \mu_0$, el estadístico:

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

tiene distribución de Student con $n-1$ grados de libertad. Usaremos entonces este estadístico de prueba, del mismo modo que antes usamos el Z.

Ejemplo 6.3

El zooplancton se compone de todos los animales oceánicos que se dejan arrastrar pasivamente por el movimiento del agua. Se extrajeron 9 muestras de agua en las proximidades de una isla, se determinó el número de individuos por m^3 y se obtuvieron los siguientes valores:

5000	5700	4450	4500	4825	4025	3700	4900	3750

Suponemos que el número de individuos por m^3 de agua está normalmente distribuido.

El investigador sospecha, por tener experiencia en el tema, que en esa zona la cantidad media de zooplancton por m^3 de agua supera los 4200 individuos y quiere verificar esa teoría. ¿Proveen los datos suficiente evidencia para apoyar esta suposición? Usar $\alpha = 0.05$.

El modelo en este ejemplo es: $X_1, X_2, \ldots, X_9$ una m.a. donde cada X_i es el número de individuos por m^3 de agua de la muestra i, con $i = 1, \ldots, 9$. Suponemos que las X_i tienen distribución $N(\mu, \sigma^2)$. Podemos enunciar el problema como:

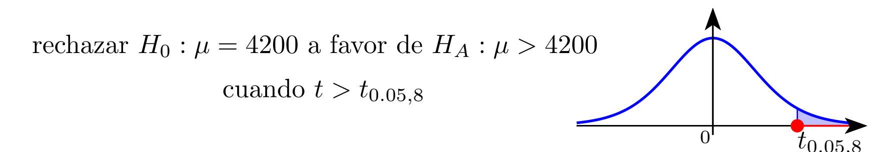
$$H_0 : \mu = 4200 \quad vs. \quad H_A : \mu > 4200$$

Como no conocemos la varianza usaremos el siguiente estadístico de prueba:

$$T = \frac{\overline{X} - 4200}{S/\sqrt{9}}$$

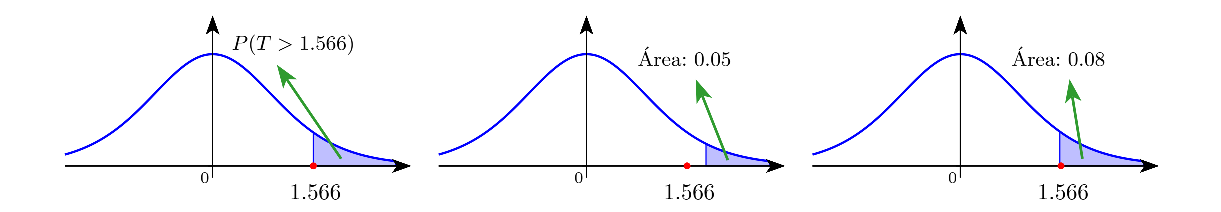
que, bajo el modelo supuesto y cuando la hipótesis nula H_0 es verdadera, tiene distribución de Student con $n-1 = 8$ grados de libertad. Luego, se puede definir la regla de decisión del siguiente modo:

$$\text{rechazar } H_0 : \mu = 4200 \text{ a favor de } H_A : \mu > 4200$$
$$\text{cuando } t > t_{0.05,8}$$



Se busca en la Tabla de Student para 8 grados de libertad, resulta $t_{0.05,8} = 1.8595$. Esto significa que la región de rechazo está a la derecha de 1.86. Reemplazando por los valores de la media muestral $\overline{x} = 4538.9$ y la desviación típica muestral $s = 649.29$, calculamos el valor del estadístico y obtenemos $t = 3(4538.9 - 4200)/649.29 = 1.566$. Este valor no cae en la región de rechazo, entonces no podemos afirmar a nivel 0.05 que el número medio de individuos de la población de zooplancton sea mayor que 4200 por m^3.

Como no tenemos una tabla que nos permita calcular la probabilidad $P(T > 1.566)$, no podemos calcular exactamente el valor-p, pero teniendo en cuenta que es el mínimo nivel de significación con el que rechazamos H_0 con los datos observados, ya podemos afirmar que valor-$p > 0.05$ (ya que con nivel $\alpha = 0.05$ no pudimos rechazar H_0), entonces vemos si es posible rechazar H_0 con un nivel menos exigente. Podemos ver en la Tabla de Student con 8 grados de libertad que el valor crítico correspondiente a $\alpha = 0.08$ es 1.5489 esto significa que con nivel 0.08 podemos rechazar H_0, entonces el valor-$p < 0.08$. Resumiendo $0.05 <$ valor-$p < 0.08$, como se muestra en la siguiente gráfica.



Ejemplo 6.4

Volviendo a la situación de determinar si un método de medición tiene error sistemático, consideremos un ejemplo más realista donde no se conoce el desvío estándar. Se hacen 10 determinaciones del contenido de níquel para una aleación estándar preparada de modo que se conoce el verdadero valor del contenido, que es 4.44 %. Se obtienen los siguientes valores:

4.32 4.31 4.50 4.12 4.43 4.36 4.48 4.28 4.18 4.42

La pregunta que nos formulamos es: ¿con estos 10 datos podemos afirmar que el método de medición tiene error sistemático?

Tenemos una m.a. $X_1, X_2, \ldots, X_{10}$, donde X_i es la i-ésima determinación del contenido de

níquel en la aleación, y suponemos que las X_i tiene distribución $N(\mu, \sigma^2)$, si el método no tiene error sistemático, la media es igual al verdadero valor. Las hipótesis a contrastar en este caso son:

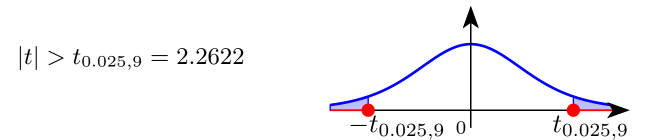
$$H_0 : \mu = 4.44 \quad vs. \quad H_A : \mu \neq 4.44,$$

el estadístico de prueba es:

$$T = \frac{\overline{X} - 4.44}{S/\sqrt{10}}$$

que tiene distribución de Student con 9 grados de libertad cuando H_0 es verdadera. Luego, la zona de rechazo con nivel $\alpha = 0.05$ es:

$$|t| > t_{0.025,9} = 2.2622$$



Haciendo los cálculos se obtiene: $\overline{x} = 4.34$ y $s = 0.1243$ y reemplazando en la fórmula del estadístico de prueba, se obtiene: $t = -2.5441$. Vemos que el valor del estadístico de prueba cae en la región de rechazo, y en consecuencia, podemos afirmar con nivel $\alpha = 0.05$ que el método tiene error sistemático.

Como hemos rechazado H_0 con $\alpha = 0.05$, sabemos que valor-$p < 0.05$. Podemos ver si aún rechazamos con un nivel más exigente. Si eligiéramos $\alpha = 0.02$, el valor crítico es $t_{0.01,9} = 2.8214$ y por lo tanto el valor observado $t = -2.5441$ no cae en la correspondiente región de rechazo, entonces valor-$p > 0.02$. Es decir:

$$0.02 < \text{ valor-}p < 0.05$$

■

Podemos resumir los diferentes tests para la media μ de la siguiente manera:

RESUMEN 6.2

Para la m.a. $X_1, X_2, \ldots, X_n$ con distribución normal con σ desconocido.			
Hipótesis nula	$H_0 : \mu = \mu_0$		
Valor del estadístico de prueba	$t = \dfrac{\sqrt{n}\,(\overline{x} - \mu_0)}{s}$		
Hipótesis alternativa	$H_A : \mu < \mu_0$	$H_A : \mu > \mu_0$	$H_A : \mu \neq \mu_0$
Región de rechazo con nivel α	$t < -t_{\alpha,n-1}$	$t > t_{\alpha,n-1}$	$\lvert t\rvert > t_{\alpha/2,n-1}$

EJERCICIO 6.2

Los siguientes datos son medias de consumo de oxígeno (en ml) durante la incubación de una m.a. de 14 suspensiones celulares:

14.0 14.1 14.5 13.2 11.2 14.1 12.2 11.1 13.7 13.2 16.0 12.8 14.4 12.9 ¿Proporcionan estos datos suficiente evidencia, en un nivel de significación de 0.05, de que la media del consumo de oxígeno es menor a 14 ml?, ¿qué supuestos se deben cumplir?

Test de hipótesis para la media de una distribución desconocida (muestras grandes)

Del mismo modo que en la construcción de un intervalo de confianza, cuando no conocemos la distribución de los datos, si la muestra es suficientemente grande, podemos utilizar el Teorema Central del Límite.

Ejemplo 6.5

Recordemos el Ejemplo 5.14. ¿Se puede afirmar, en base a estos datos, que los peces de esa región tienen niveles medios de zinc mayores a 8.2 $\mu g/g$?

Aquí tenemos una m. a. $X_1, X_2, \ldots, X_{56}$ donde cada X_i es la concentración de zinc en el hígado ($\mu g/g$) del i-ésimo pez examinado y desconocemos su distribución, pero suponemos que tiene media $\mu = E(X_i)$ y varianza $\sigma^2 = V(X_i)$.

El problema puede plantearse como:

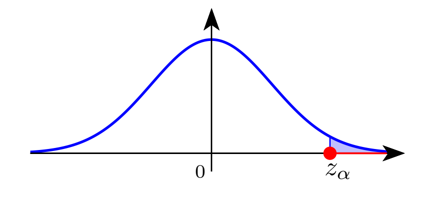
$$H_0 : \mu = 8.2 \quad vs. \quad H_A : \mu > 8.2$$

y en este caso, como n es grande, podemos aplicar el resultado del Teorema Central del Límite y usar el estadístico de prueba:

$$Z = \sqrt{56}\,\frac{(\overline{X} - 8.2)}{S}$$

ya que, según este teorema, cuando $\mu = 8.2$ tiene una distribución aproximadamente $N(0, 1)$. Entonces podemos definir, como siempre, una regla de decisión:

$$\text{rechazar } H_0 : \mu = 8.2 \text{ a favor de } H_A : \mu > 8.2 \text{ cuando } z > z_\alpha$$



con los datos del ejemplo, $\overline{x} = 9.15\ \mu g/g$ y $s = 1.27\ \mu g/g$, reemplazando en el estadístico, obtenemos un valor $z = \sqrt{56}\ (9.15 - 8.2)/1.27 = 5.598$, vemos en la Tabla de la distribución Normal que el valor-$p = P(Z > 5.598) < 0.0001$, esto significa que hay muy fuerte evidencia para rechazar H_0, y se puede rechazar con cualquier nivel de significación razonable. O sea, podemos concluir que la concentración media de zinc en el hígado de los peces de esa región es superior a $8.2\ \mu g/g$.

■

Podemos resumir los casos de tests para la media de una distribución desconocida, cuando n es grande, de la siguiente manera:

RESUMEN 6.3

Para la m.a. $X_1, X_2, \ldots, X_n$ con distribución desconocida y n grande.			
Hipótesis nula	$H_0 : \mu = \mu_0$		
Valor del estadístico de prueba	$z = \dfrac{\sqrt{n}\,(\overline{x} - \mu_0)}{s}$		
Hipótesis alternativa	$H_A : \mu < \mu_0$	$H_A : \mu > \mu_0$	$H_A : \mu \neq \mu_0$
Región de rechazo con nivel aproximado α	$z < -z_\alpha$	$z > z_\alpha$	$\lvert z \rvert > z_{\alpha/2}$



Observación:

En este caso el nivel es aproximado, porque no conocemos la distribución exacta del estadístico de prueba y estamos utilizando una aproximación.

EJERCICIO 6.3

En una muestra de 49 adolescentes que se prestaron como sujetos para un estudio inmunológico, una variable de interés fue la prueba del diámetro de reacción de la piel a un antígeno. La media y la desviación estándar de la muestra fueron de 21 y 11 mm, respectivamente. ¿Es posible concluir a partir de estos datos que la media poblacional es diferente a 30? Tomar $\alpha = 0.01$.

Test de hipótesis para una proporción

Ya vimos al tratar el tema intervalo de confianza para una proporción, en el Capítulo 5, que si tenemos una v.a. X con distribución $B(n, p)$ entonces, si n es suficientemente grande, la distribución de:

$$Z = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

es aproximadamente $N(0, 1)$ por el Teorema Central del Límite.

Ejemplo 6.6

Una de las célebres leyes de Murphy establece que si se deja caer al suelo una tostada untada con dulce, la probabilidad de que caiga del lado del dulce es mayor que la de que caiga del lado del pan. Para verificarla, se realizó un experimento: se dejaron caer 1000 tostadas untadas con mermelada, de las cuales 540 cayeron del lado del dulce. ¿Qué podría concluir a nivel $\alpha = 0.05$?

Para modelizar este problema definimos X= "el número de tostadas que caen del lado del dulce, entre las 1000 del experimento". Esta v.a. tiene distribución $B(1000, p)$, según la ley de Murphy este valor de p es mayor que 0.5. Entonces, para verificar esta ley planteamos el siguiente test:

$$H_0 : p = 0.5 \quad vs. \quad H_A : p > 0.5$$

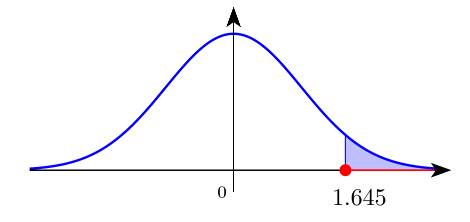
Si H_0 es verdadera ($p = 0.5$), según el TCL,

$$Z = \frac{\widehat{p} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{1000}}}$$

tiene distribución aproximadamente $N(0, 1)$, este será nuestro estadístico de prueba.

Si usamos un nivel $\alpha = 0.05$, la regla de decisión será:

$$\text{rechazar } H_0 : p = 0.5 \text{ a favor de } H_A : p > 0.5 \text{ cuando } z > z_{0.05} = 1.645$$



En el experimento se obtuvo $\widehat{p} = 0.54$, reemplazando en el estadístico de prueba se obtiene $z = 2.53$, que cae en la zona de rechazo. Conclusión, en base a este experimento, podemos afirmar (con nivel 0.05) que la ley de Murphy es verdadera.

Si queremos conocer el valor-p, calculamos $P(Z > 2.53) \cong 1 - \Phi(2.53) = 0.0057$. Esto significa que podemos afirmar que esta ley de Muphy es verdadera con cualquier nivel ≥ 0.0057.

Podemos resumir los casos de tests para una proporción como sigue:

RESUMEN 6.4

Para la v.a. X con distribución $B(n,p)$ y n grande.			
Hipótesis nula	$H_0 : p = p_0$		
Valor del estadístico de prueba	$z = \dfrac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$		
Hipótesis alternativa	$H_A : p < p_0$	$H_A : p > p_0$	$H_A : p \neq p_0$
Región de rechazo con nivel aproximado α	$z < -z_\alpha$	$z > z_\alpha$	$\lvert z \rvert > z_{\alpha/2}$



Observación:

En este caso también el nivel es aproximado, porque no conocemos la distribución exacta del estadístico de prueba y estamos utilizando una aproximación.

EJERCICIO 6.4

1. En un estudio se encontró que 66 % de los niños en una muestra de 670 completaron toda la serie de vacunas contra la hepatitis B. ¿Es posible concluir que, con base a estos datos, en la población muestreada, más del 60 % tienen la serie completa de vacunas contra la hepatitis B? Tomar $\alpha = 0.01$.

2. En una investigación de consumidores de drogas intravenosas en una ciudad grande, encontraron a 18 de 423 individuos con HIV positivo. Se pretende saber si es posible concluir que los consumidores de drogas intravenosas en la población muestreada tienen HIV positivo en un porcentaje diferente del 5 %.

Test de hipótesis para la varianza de una distribución normal

Cuando hicimos intervalos de confianza para la varianza ya vimos que, para la m.a. $X_1, X_2, \ldots, X_n$ con distribución normal, la función:

$$V = \frac{(n-1)\,S^2}{\sigma^2}$$

tiene distribución Chi-cuadrado con $n-1$ grados de libertad. Si tenemos una distribución normal y queremos hacer un test para la varianza, usaremos como estadístico de prueba esa función reemplazando σ^2 por el valor de la hipótesis nula.

Ejemplo 6.7

Consideremos el ejemplo del método de medición del contenido de níquel, Ejemplo 6.4. Si se supone que un valor aceptable de σ debe ser menor que 0.3, ¿podemos afirmar que el método es aceptable con $\alpha = 0.05$?

Ya dijimos que las 10 determinaciones son v.a. con distribución $N(\mu, \sigma^2)$. El problema puede plantearse como:

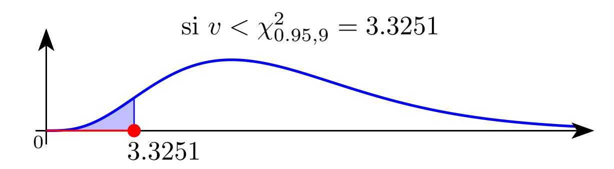
$$H_0 : \sigma^2 = 0.3^2 \quad vs. \quad H_A : \sigma^2 < 0.3^2$$

El estadístico de prueba será:

$$V = \frac{9 \times S^2}{0.3^2}$$

que tiene distribución Chi-cuadrado con 9 grados de libertad si H_0 es verdadera. La regla de decisión es:

$$\text{rechazar } H_0 : \sigma^2 = 0.3^2 \text{ a favor de } H_A : \sigma^2 < 0.3^2$$
$$\text{si } v < \chi^2_{0.95,9} = 3.3251$$



En este caso, el valor del estadístico de prueba es $v = 1.5451$ que cae en la zona de rechazo de nivel 0.05. Esto significa que podemos afirmar, con nivel 0.05, que el desvío estándar del error de medición es aceptable.

■

Podemos resumir los casos de tests para la varianza de una distribución normal como sigue:

RESUMEN 6.5

Para la m.a. $X_1, X_2, \ldots, X_n$ con distribución normal con σ desconocido.			
Hipótesis nula	$H_0 : \sigma^2 = \sigma_0^2$		
Valor del estadístico de prueba	$v = \frac{(n-1)\, s^2}{\sigma_0^2}$		
Hipótesis alternativa	$H_A : \sigma^2 < \sigma_0^2$	$H_A : \sigma^2 > \sigma_0^2$	$H_A : \sigma^2 \neq \sigma_0^2$
Región de rechazo con nivel α	$v < \chi^2_{1-\alpha,n-1}$	$v > \chi^2_{\alpha,n-1}$	$v > \chi^2_{\alpha/2,n-1}$ ó $v < \chi^2_{1-\alpha/2,n-1}$

EJERCICIO 6.5

1. Se registraron los valores de la capacidad vital de una muestra de 10 pacientes con obstrucción crónica severa de las vías respiratorias. La varianza de las 10 observaciones fue de 0.75. Pruebe la hipótesis nula que indica que la varianza de la población es de 1. Con $\alpha = 0.05$. ¿Qué supuestos deben hacerse?

2. En un estudio realizado en 15 pacientes con enfermedad sarcoide pulmonar, se midieron las concentraciones de gases en la sangre. La varianza de los valores de PaO_2 (en $mmHg$) fue de 450. Pruebe la hipótesis alternativa según la cual la varianza de la población es mayor que 250. Con $\alpha = 0.05$.

Referencias

Agresti, A. & Franklin, C. A. (2009). *Statistics: The Art and Science of learning from Data*. Pearson New International edition.

Altman, D. G. (1990). *Practical Statistics for Medical Research*. Published by Chapman & Hall.

Daniel, W. (2002). *Bioestadística: Base para el análisis de las ciencias de la salud*. Ed. Limusa Wiley.

Devore Jay, L. (2001). *Probabilidad y Estadística para Ingeniería y Ciencias*. Ed. Books/Cole Publishing Company.

Dixon, W. & Massey, F. (1970). *Introducción al Análisis Estadístico*. México. Libros Mc Graw-Hill.

Maronna, R. (1995). *Probabilidad y Estadística Elementales para Estudiantes de Ciencias*. Buenos Aires.Ed. Exactas.

Mendenhall, W., Beaver, R. J. & Beaver, B. M. (2006). *Introducción a la Probabilidad y Estadística*. México. Cengage Learning Editores.

Ross, S. M. (1987). *Introduction to Probability and Statistics for Enginees and Scientists*. Published by John Wiley & Sons.

Wackerly, D. D., Mendenhall, W. & Scheaffer, R. L. (2010). *Estadística Matemática con aplicaciones*. México. Cengage Learning Editores.

Walpole, R. E. & Myers, R. H. (2007). *Probabilidad y Estadística para Ingeniería y Ciencias*. México. Ediciones McGraw-Hill.

CAPÍTULO 7

Inferencias basadas en dos muestras

Introducción

En el capítulo anterior vimos tests para comparar la media de una población con un valor fijo μ_0. Sin embargo, en la mayoría de las aplicaciones, interesa comparar dos poblaciones. Por ejemplo, para evaluar el efecto de un nuevo tratamiento, se suele comparar un grupo de individuos al que se aplica el nuevo tratamiento con otro grupo al que se le aplica otro tratamiento o un placebo; en otros casos se comparan individuos expuestos a un factor de riesgo con otros que no lo están, o se desea comparar los resultados de dos métodos de medición, etc.

Los procedimientos para construir intervalos de confianza para la diferencia de medias y realizar tests para comparación de medias, son similares a los que vimos antes. Lo principal es encontrar la función pivote y el estadístico de prueba adecuado para cada situación.

Intervalo de confianza y test para la diferencia de medias de dos poblaciones normales con varianzas conocidas

Sean $X_1, X_2, \ldots, X_{n_1}$ una m.a. de una distribución $N(\mu_1, \sigma_1^2)$ e $Y_1, Y_2, \ldots, Y_{n_2}$ una m.a. de una distribución $N(\mu_2, \sigma_2^2)$, independientes entre sí y con varianzas conocidas.

Un estimador insesgado para la diferencias de las medias, $\mu_1 - \mu_2$, es $\overline{X} - \overline{Y}$ y sabemos que $\overline{X} - \overline{Y}$ tiene distribución $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$, entonces si deseamos construir un intervalo

de confianza para $\mu_1 - \mu_2$, la función pivote será:

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

que tiene distribución $N(0,1)$. Y el intervalo de confianza para $\mu_1 - \mu_2$ resultará:

$$IC_\alpha(\mu_1 - \mu_2) = \left(\overline{X} - \overline{Y} - z_{\alpha/2} \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \overline{X} - \overline{Y} + z_{\alpha/2} \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

Ejemplo 7.1

A un equipo de investigación le interesa conocer la diferencia entre las concentraciones de ácido úrico en pacientes con y sin una determinada enfermedad. En un hospital especializado para el tratamiento de esa enfermedad, una muestra de 12 pacientes presenta una media de 4.5 $mg/100\ ml$. En un hospital general se encontró que una muestra de 15 individuos sin esta enfermedad de la misma edad y sexo presenta un nivel medio de 3.4 $mg/100\ ml$. Si es razonable suponer que las dos poblaciones de valores muestran una distribución normal y sus varianzas son iguales a 1 y 1.5, respectivamente, calculemos el intervalo de confianza de 95 % para la diferencias de medias.

Primero modelicemos el problema. Sea $X_1, X_2, \ldots, X_{12}$ una m.a. con cada $X_i \sim N(\mu_1, 1)$, que indica el nivel de ácido úrico en el paciente i (medido en $mg/100\ ml$) con la enfermedad. Además tenemos $Y_1, Y_2, \ldots, Y_{15}$ una m.a. con cada $Y_j \sim N(\mu_2, 1.5)$, que indica el nivel de ácido úrico en el paciente j (medido en $mg/100\ ml$) sin la enfermedad. Las v.a. X_i y Y_j son independientes.

Entonces tenemos que $\overline{x} = 4.5$, $\overline{y} = 3.4$ y $z_{0.025} = 1.96$, así el intervalo que deseamos es:

$$IC_{0.95}(\mu_1 - \mu_2) = 4.5 - 3.4 \pm 1.96 \times \sqrt{\frac{1}{12} + \frac{1.5}{15}} = (0.2608, 1.9392)$$

Esto significa que tenemos el 95 % de confianza de que la diferencia real $\mu_1 - \mu_2$, está entre 0.2608 y 1.9392. Recordemos la interpretación de intervalo de confianza vista en el Capítulo 5.

■

Si deseamos contrastar hipótesis sobre $\mu_1 - \mu_2$, donde $H_0 : \mu_1 - \mu_2 = \Delta_0$ debemos usar el estadístico de prueba:

$$\frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

que tiene distribución $N(0,1)$ cuando H_0 es verdadera. Entonces el resumen para los tests de hipótesis para $\mu_1 - \mu_2$ será:

RESUMEN 7.1

Para las m.a. $X_1, X_2, \ldots, X_{n_1}$ con distribución $N(\mu_1, \sigma_1^2)$ e $Y_1, Y_2, \ldots, Y_{n_2}$ con distribución $N(\mu_2, \sigma_2^2)$, independientes entre sí.			
Hipótesis nula	$H_0 : \mu_1 - \mu_2 = \Delta_0$		
Valor del estadístico de prueba	$z = \dfrac{\overline{x} - \overline{y} - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$		
Hipótesis alternativa	$H_A : \mu_1 - \mu_2 < \Delta_0$	$H_A : \mu_1 - \mu_2 > \Delta_0$	$H_A : \mu_1 - \mu_2 \neq \Delta_0$
Región de rechazo con nivel α	$z < -z_\alpha$	$z > z_\alpha$	$\lvert z\rvert > z_{\alpha/2}$

En muchas situaciones sólo interesa saber si las medias de las dos poblaciones son diferentes, en ese caso $\Delta_0 = 0$.

Ejemplo 7.2

Se realizó un estudio para determinar la resistencia a la ruptura de dos tipos de acero. Para una muestra aleatoria formada por 20 especímenes de acero laminado en frío la resistencia promedio muestral fue $\overline{x} = 29.8\ ksi$. Al estudiar una segunda muestra aleatoria de 25 especímenes de acero galvanizado de dos lados se obtuvo una resistencia promedio muestral $\overline{y} = 32.7\ ksi$. Se supone que las distribuciones de la resistencia a la ruptura de los dos tipos de acero son normales con $\sigma_1 = 4$, para el acero laminado en frío, y $\sigma_2 = 5$ para el acero galvanizado. ¿Indican los datos que las medias de resistencia a la ruptura son diferentes para los dos tipos de acero?

Modelizando: tenemos $X_1, X_2, \ldots, X_{20}$ una m.a. de una $N(\mu_1, 16)$ e $Y_1, Y_2, \ldots, Y_{25}$ una m.a. de una distribución $N(\mu_2, 25)$, independientes entre sí. Cada X_i indica la resistencia a la ruptura del i-ésimo espécimen de acero laminado en frío y cada Y_j la resistencia a la ruptura del j-ésimo espécimen de acero galvanizado. En este caso el problema se plantea como:

$$H_0 : \mu_1 = \mu_2 \quad vs. \quad H_A : \mu_1 \neq \mu_2$$

o en forma equivalente

$$H_0 : \mu_1 - \mu_2 = 0 \quad vs. \quad H_A : \mu_1 - \mu_2 \neq 0$$

Para este problema el estadístico de prueba será:

$$Z = \frac{\overline{X} - \overline{Y} - 0}{\sqrt{\frac{16}{20} + \frac{25}{25}}}$$

que tiene distribución $N(0, 1)$, cuando H_0 es verdadera, y la regla de decisión es:

rechazar H_0 si el valor $|z| = \dfrac{|\overline{x} - \overline{y}|}{\sqrt{\frac{16}{20} + \frac{25}{25}}} > z_{\alpha/2}$

Si elegimos un nivel de significación $\alpha = 0.05$, el punto crítico será $z_{\alpha/2} = 1.96$. Entonces, reemplazando con los valores muestrales vemos que en este caso el valor del estadístico de prueba es: $z = -2.16$, que cae en la zona de rechazo. Por lo tanto, podemos concluir que para un nivel 0.05, afirmamos que las medias de resistencia a la ruptura difieren para los dos tipos de acero.

Si calculamos el valor-p:

$$\text{valor-}p = P\big(|Z| > |-2.16|\big) = 1 - P\big(|Z| \leq 2.16\big) = 1 - P\big(-2.16 \leq Z \leq 2.16\big)$$

$$= 1 - \big[\Phi(2.16) - \Phi(-2.16)\big] = 2 \times (1 - 0.9846) = 0.0308$$

esto significa que, hasta con un nivel 0.0308, podemos afirmar que la resistencia a la ruptura de los dos tipos de acero es diferente.

■

EJERCICIO 7.1

Veinticuatro animales de laboratorio con deficiencia de vitamina D fueron divididos en 2 grupos iguales. El grupo 1 recibió un tratamiento consistente en una dieta que proporcionaba vitamina D. Al segundo grupo no se le administró tratamiento. Al término del periodo experimental, se midieron las concentraciones de calcio en suero, obteniéndose los siguientes resultados:

Grupo	Media muestral
Con tratamiento	11.1 $mg/100\ ml$
Sin tratamiento	7.8 $mg/100\ ml$

Considerar que las poblaciones siguen una distribución normal con $\sigma_1 = 1.5$ y $\sigma_2 = 2$.

1. Construir el intervalo de confianza para la diferencia de medias con nivel del 98 %.

2. ¿Sugieren los datos que la concentración media de calcio en suero en el grupo tratado es superior en más de 2 $mg/100\ ml$ con respecto al grupo sin tratar? Formule y pruebe las hipótesis apropiadas utilizando un nivel de significación de 0.05.

Pero no siempre que tenemos dos muestras aleatorias de distribuciones normales, conocemos sus varianzas. Veamos ahora este caso.

Intervalo de confianza y test para la diferencia de medias de dos poblaciones normales con varianzas desconocidas pero iguales

Sean dos m.a. $X_1, X_2, \ldots, X_{n_1}$ con distribución $N(\mu_1, \sigma^2)$ e $Y_1, Y_2, \ldots, Y_{n_2}$ con distribución $N(\mu_2, \sigma^2)$ e independientes entre sí, con varianzas desconocidas pero iguales.

Sabemos que $\overline{X} - \overline{Y}$ es un estimador razonable para la diferencias de las medias, $\mu_1 - \mu_2$, y cuando las X_i y las Y_j tienen distribución normal y son muestras independientes, $\overline{X} - \overline{Y}$ tiene distribución normal con

$$E(\overline{X} - \overline{Y}) = \mu_1 - \mu_2 \quad \text{y} \quad Var(\overline{X} - \overline{Y}) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

entonces la función:

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

tiene distribución $N(0, 1)$, pero dado que no conocemos σ debemos reemplazarlo por un estimador. Para este caso usaremos el **estimador ponderado de la varianza** que se define como:

$$S_p^2 = \frac{\sum_{i=1}^{n_1} (X_i - \overline{X})^2 + \sum_{i=1}^{n_2} (Y_i - \overline{Y})^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{7.1}$$

que es insesgado y usaremos $S_p = \sqrt{S_p^2}$ como estimador de σ. Luego, si reemplazamos σ por S_p, obtenemos la función pivote:

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

que tiene distribución de Student con $n_1 + n_2 - 2$ grados de libertad. Luego el intervalo de confianza de nivel $1 - \alpha$ es:

$$IC_{1-\alpha}(\mu_1 - \mu_2) = \overline{X} - \overline{Y} \pm t_{\alpha/2, n_1+n_2-2} \times S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Ejemplo 7.3

El objetivo de un estudio de aptitud física fue determinar los resultados del ejercicio por un tiempo prolongado en los ejecutivos de una compañía inscriptos en un programa supervisado de acondicionamiento físico. Se registraron datos de 13 individuos (el grupo deportista) que voluntariamente se inscribieron al programa y que permanecieron activos por 13 años

en promedio, y de 17 individuos (el segundo grupo, el sedentario) que decidieron no inscribirse. Entre los datos que se registraron acerca de los individuos está el número máximo de sentadillas realizadas en 30 segundos. El grupo deportista obtuvo una media y una desviación estándar de 21 y 4.9, respectivamente. La media y desviación estándar para el grupo sedentario fueron 12.1 y 5.6 respectivamente. Se considera que las dos poblaciones de mediciones de acondicionamiento muscular siguen una distribución aproximadamente normal, y que las varianzas para ambas poblaciones son iguales. Se pretende elaborar un intervalo del 98 % para la diferencias entre las medias de las poblaciones representadas por las dos muestras.

Primero modelicemos el problema. Tenemos dos grupos de m.a.: el primero son $X_1, X_2, \ldots, X_{13}$ con cada $X_i \sim N(\mu_1, \sigma_1^2)$, que indica el número máximo de sentadillas realizadas en 30 segundos por el individuo i del grupo deportista; y el segundo son $Y_1, Y_2, \ldots, Y_{17}$ con cada $Y_j \sim N(\mu_2, \sigma_2^2)$, que indica el número máximo de sentadillas realizadas en 30 segundos por el individuo j del grupo sedentario. Las v.a. X_i e Y_j son independientes.

Entonces tenemos que $\overline{x} = 21$, $\overline{y} = 12.1$, $s_1^2 = 4.9^2$ y $s_2^2 = 5.6^2$. Calculemos el valor de S_p^2 observado:

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(13-1)\times 4.9^2 + (17-1)\times 5.6^2}{13+17-2} = 28.21$$

es decir, $s_p = 5.3113$. Consultando la Tabla de la distribución Student con 13 + 17 - 2 = 28 grados de libertad con un nivel de confianza de 98 %, se observa que el valor crítico es 2.4671. Por lo tanto, el intervalo de confianza de 98 % para la diferencia de las medias poblacionales se obtiene de la siguiente manera:

$$IC_{0.98}(\mu_1-\mu_2) = 21 - 12.1 \pm 2.4671 \times 5.3113\sqrt{\frac{1}{13}+\frac{1}{17}} = (4.0722, 13.7278)$$

Se tiene una confianza del 98 % de que la diferencia entre las medias de las poblaciones están entre 4.0722 y 13.7278.

■

Si deseamos realizar algún test sobre $\mu_1 - \mu_2$, el estadístico de prueba será:

$$T = \frac{\overline{X} - \overline{Y} - \Delta_0}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

que bajo la hipótesis nula, $H_0 : \mu_1 - \mu_2 = \Delta_0$, tiene distribución de Student con $n_1 + n_2 - 2$ grados de libertad. Resumimos entonces los tests de comparación de medias:

RESUMEN 7.2

Para las m.a. $X_1, X_2, \ldots, X_{n_1}$ con distribución $N(\mu_1, \sigma^2)$ e $Y_1, Y_2, \ldots, Y_{n_2}$ con distribución $N(\mu_2, \sigma^2)$, independientes entre sí.			
Hipótesis nula	$H_0 : \mu_1 - \mu_2 = \Delta_0$		
Valor del estadístico de prueba	$t = \dfrac{\overline{x} - \overline{y} - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$		
Hipótesis alternativa	$H_A : \mu_1 - \mu_2 < \Delta_0$	$H_A : \mu_1 - \mu_2 > \Delta_0$	$H_A : \mu_1 - \mu_2 \neq \Delta_0$
Región de rechazo con nivel α	$t < -t_{\alpha,n_1+n_2-2}$	$t > t_{\alpha,n_1+n_2-2}$	$\lvert t\rvert > t_{\alpha/2,n_1+n_2-2}$

Ejemplo 7.4

Se desea verificar si hay diferencia en el nivel medio de hierro en sangre, entre los niños con fibrosis quística y los niños sanos. Se seleccionaron 9 niños con fibrosis quística y se les realizaron las mediciones de hierro en sangre, obteniendose $\overline{x} = 11.9\ \mu mol/l$ y $s_1 = 6.3\ \mu mol/l$. También se realizaron mediciones de hierro en sangre sobre de una muestra de 13 niños sanos, obteniendose $\overline{y} = 18.9\ \mu mol/l$ y $s_2 = 5.9\ \mu mol/l$. Asumimos que la distribución de los valores de hierro en ambas poblaciones es normal con igual varianza.

Modelizando: tenemos $X_1, X_2, \ldots, X_9$ una m.a. con cada $X_i \sim N(\mu_1, \sigma^2)$, que indica el nivel de hierro en sangre del niño i con fibrosis quística (medido en $\mu mol/l$). Además tenemos $Y_1, Y_2, \ldots, Y_{13}$ una m.a. con cada $Y_j \sim N(\mu_2, \sigma^2)$, que indica el nivel de hierro en sangre del niño j sano (medido en $\mu mol/l$). Las v.a. X_i e Y_j son independientes.

En este caso el problema se plantea como:

$$H_0 : \mu_1 = \mu_2 \quad vs. \quad H_A : \mu_1 \neq \mu_2$$

o en forma equivalente

$$H_0 : \mu_1 - \mu_2 = 0 \quad vs. \quad H_A : \mu_1 - \mu_2 \neq 0$$

La regla de decisión, para este test bilateral, será:

rechazar $H_0 : \mu_1 = \mu_2$ a favor de $H_A : \mu_1 \neq \mu_2$ si el valor $|t| = \dfrac{|\overline{x} - \overline{y}|}{s_p \sqrt{\frac{1}{9} + \frac{1}{13}}} > t_{\alpha/2,20}$

Si deseamos un nivel de significación $\alpha = 0.05$, el valor crítico será $t_{0.025,20} = 2.086$. Aquí, tenemos los valores $s_p^2 = \frac{8 \times 6.3^2 + 12 \times 5.9^2}{20} = 36.762$ y el valor del estadístico de prueba es $t = \frac{|11.9 - 18.9|}{6.0632\sqrt{1/9 + 1/13}} = 2.6624$. Como el valor del estadístico de prueba cae en la zona de rechazo, se puede rechazar la hipótesis nula con nivel $\alpha = 0.05$. Entonces podemos afirmar, con

$\alpha = 0.05$, que el nivel medio de hierro en sangre de los niños con fibrosis quística es diferente al de los niños sanos.

También podemos ver que el valor crítico $t_{0.01,20} = 2.528$ (para el test bilateral correspondiente a $\alpha = 0.02$) y el $t_{0.005,20} = 2.8453$ (que corresponde a $\alpha = 0.01$), esto significa que podemos rechazar H_0 con nivel $\alpha = 0.02$, pero no con nivel $\alpha = 0.01$. Se suele decir que el resultado es significativo a nivel 0.02, o que el valor-p estaría entre 0.01 y 0.02.

■

EJERCICIO 7.2

1. Demostrar que S_p^2, definido en (7.1), es un estimador insesgado para σ^2 cuando las v.a. $X_1, X_2, \ldots, X_{n_1}$ tienen distribución $N(\mu_1, \sigma^2)$ y las v.a. $Y_1, Y_2, \ldots, Y_{n_2}$ tienen distribución $N(\mu_2, \sigma^2)$ con varianzas desconocidas pero iguales y son independientes entre sí.

2. A dos grupos de adultos se les hicieron pruebas de audiometría. El primer grupo estuvo formado por 11 adultos que trabajan temporalmente en carpinterías. La calificación media para este grupo fue de 26 decibelios (dB) con una desviación estándar de 5 dB. El segundo grupo, que incluyó a 14 personas que trabajan temporalmente en peluquerías, tuvo una calificación promedio de 21 dB con una desviación estándar de 6 dB. Suponga que las poblaciones siguen una distribución normal con varianzas iguales, ¿hay evidencia suficiente para afirmar que los niveles medios de la audiometría para los que han trabajado temporalmente en las carpintería es mayor a 3 dB que los que trabajan temporalmente en peluquerías?

En muchas aplicaciones, la suposición de que las varianzas de las dos poblaciones son iguales es poco realista.

Intervalo de confianza y test para la diferencia de medias de dos poblaciones normales con varianzas desconocidas

Sean dos m.a. $X_1, X_2, \ldots, X_{n_1}$ con distribución $N(\mu_1, \sigma_1^2)$ e $Y_1, Y_2, \ldots, Y_{n_2}$ con distribución $N(\mu_2, \sigma_2^2)$ independientes entre sí y con varianzas desconocidas. Luego la función

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

tiene distribución $N(0, 1)$, pero como desconocemos las varianzas, la función pivote será:

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

la distribución de esta función pivote se aproxima a una Student con ν grados de libertad, donde:

$$\nu = \frac{\left[(s_1^2/n_1) + (s_2^2/n_2)\right]^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}} \tag{7.2}$$

Aclaración

Para los grados de libertad, ν, se toma al entero más próximo al valor calculado en (7.2).

Entonces el intervalo de confianza de nivel $1-\alpha$ para $\mu_1 - \mu_2$ resulta:

$$IC_{1-\alpha}(\mu_1 - \mu_2) = \left(\overline{X} - \overline{Y} - t_{\alpha/2,\nu} \times \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\, , \overline{X} - \overline{Y} + t_{\alpha/2,\nu} \times \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right)$$

Ejemplo 7.5

Se tienen datos de la actividad total del complemento serológico en 10 sujetos enfermos:

27.1	90.9	67.7	98.7	58.5	76.9	91.1	95.5	56.5	92.6

y en 20 sujetos aparentemente normales:

44.6	58.1	44.1	55.9	30.1	53.8	56.8	43.9	61.4	58.3
30.3	44.1	48.7	45.5	42.2	49.5	57.9	44.5	34.5	41.5

Asumimos que estas son observaciones de las v.a. $X_1, X_2, \ldots, X_{10}$ e $Y_1, Y_2, \ldots, Y_{20}$, donde X_i es la medición del complemento serológico del sujeto i del grupo de enfermos con $X_i \sim N(\mu_1, \sigma_1^2)$, e Y_j es la medición del complemento serológico del sujeto j del grupo de sanos con $Y_j \sim N(\mu_1, \sigma_1^2)$. Si deseamos construir un intervalo de confianza para la diferencias de medias, de nivel 0.98, tenemos que: $\overline{x} = 75.55$, $\overline{y} = 47.285$, $s_1 = 23.0133$, $s_2 = 9.2361$, $\nu = 10.4758$ (entonces se tomará 10 grados de libertad) y $t_{0.01,10} = 2.7638$, entonces obtenemos:

$$IC_{0.98}(\mu_1 - \mu_2) = 75.55 - 47.285 \pm 2.7638 \times \sqrt{\frac{23.0133^2}{10} + \frac{9.2361^2}{20}} = (7.3574, 49.1726)$$

■

Para realizar un test el estadístico de prueba es:

$$T = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

cuando $H_0 : \mu_1 - \mu_2 = \Delta_0$ es verdadera, la distribución de este estadístico se aproxima a una Student con ν grados de libertad definidos en (7.2). Entonces resumiendo los posibles tests tenemos:

RESUMEN 7.3

Para las m.a. $X_1, X_2, \dots, X_{n_1}$ con distribución $N(\mu_1, \sigma_1^2)$ e $Y_1, Y_2, \dots, Y_{n_2}$ con distribución $N(\mu_2, \sigma_2^2)$, independientes entre sí.			
Hipótesis nula	$H_0 : \mu_1 - \mu_2 = \Delta_0$		
Valor del estadístico de prueba	$t = \dfrac{\overline{x} - \overline{y} - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$		
Hipótesis alternativa	$H_A : \mu_1 - \mu_2 < \Delta_0$	$H_A : \mu_1 - \mu_2 > \Delta_0$	$H_A : \mu_1 - \mu_2 \neq \Delta_0$
Región de rechazo con nivel α	$t < -t_{\alpha,\nu}$	$t > t_{\alpha,\nu}$	$\lvert t\rvert > t_{\alpha/2,\nu}$

Ejemplo 7.6

Se propone un tratamiento para la artritis reumatoide, que es aplicado a una muestra de 6 pacientes, a los que se mide la concentración de tiol en sangre. Estos valores se comparan con los de 5 pacientes control tratados con placebo.

control	2.81	3.62	3.27	2.35	3.67	
tratamiento	1.95	2.10	2.05	1.92	2.56	2.30

Sabiendo que la concentración de tiol en sangre se distribuye normalmente, ¿hay suficiente evidencia para afirmar que el tratamiento reduce los valores de tiol? ($\alpha = 0.05$).

Definimos X_i al valor de concentración de tiol en sangre del paciente i del grupo control, $X_i \sim N(\mu_1, \sigma_1^2)$, con $i = 1, \dots, 5$ y definimos Y_j al valor de concentración de tiol en sangre del paciente j del grupo en tratamiento, $Y_j \sim N(\mu_2, \sigma_2^2)$, $j = 1, \dots, 6$. Ambas varianzas desconocidas. Para estas dos m.a. los valores calculados resultan ser: $\overline{x} = 3.144$, $\overline{y} = 2.1467$, $s_1 = 0.5615$ y $s_2 = 0.2433$.

Este caso de test se puede plantear como:

$$H_0 : \mu_1 = \mu_2 \quad vs. \quad H_A : \mu_1 > \mu_2$$

o en forma equivalente:

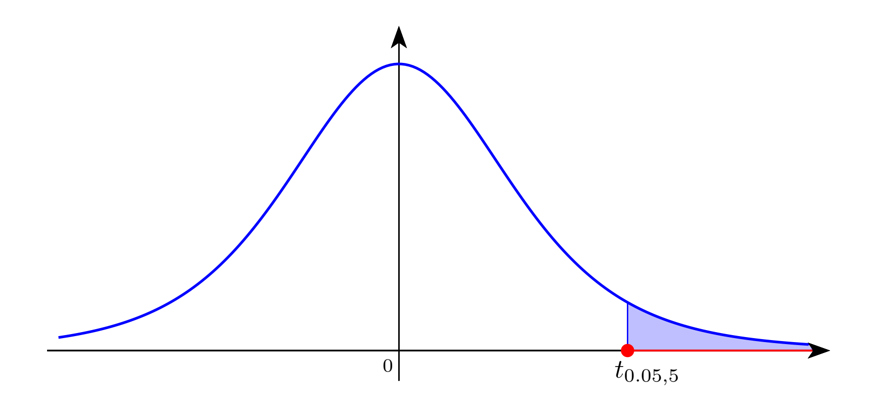
$$H_0 : \mu_1 - \mu_2 = 0 \quad vs. \quad H_A : \mu_1 - \mu_2 > 0$$

el estadístico de prueba es:

$$T = \frac{\overline{X} - \overline{Y} - 0}{\sqrt{\frac{S_1^2}{5} + \frac{S_2^2}{6}}}$$

cuando $H_0 : \mu_1 - \mu_2 = 0$ es verdadera, la distribución de este estadístico se aproxima a una Student con 5 grados de libertad (por (7.2) $\nu = 5.2468$). Luego, la regla de decisión será:

$$\text{rechazar } H_0 : \mu_1 - \mu_2 = 0 \text{ a favor de } H_A : \mu_1 - \mu_2 > 0 \text{ cuando } t > t_{0.05,5}$$



El valor del estadístico de prueba es $t = 3.6931$ y $t_{0.05,7} = 2.015$, entonces se puede rechazar la hipótesis nula a nivel 0.05. Es decir, se puede afirmar, con nivel $\alpha = 0.05$, que el tratamiento reduce los valores de tiol.

■

EJERCICIO 7.3

Los siguientes datos proporcionan información resumida sobre permeabilidad al aire, medido en $cm^3/cm^2/seg$, de dos tipos de tela: algodón y triacetato.

Tipo de tela	Tamaño muestral	media muestral	desviación estándar muestral
algodón	10	51.71	0.79
triacetato	10	136.14	3.59

Suponemos que la permeabilidad de ambos tipos de telas son normales.

1. Calcular el intervalo de confianza de nivel 95 % para la diferencia de permeabilidad media de ambos tipos de telas.

2. ¿Se puede afirmar que la permeabilidad media del triacetato supera a la permeabilidad media del algodón en más de 80 $cm^3/cm^2/seg$?

Análisis de datos apareados

La característica fundamental de las muestras apareadas, es que a cada observación en el primer grupo, le corresponde una en el segundo grupo. Generalmente se trata de dos mediciones realizadas a un mismo individuo en dos ocasiones, un ejemplo común es el experimento "antes y después", donde a cada individuo se le realiza un examen antes de aplicar un tratamiento y se vuelve a realizar ese examen después del tratamiento. En otras ocasiones el investigador relaciona cada individuo

de un grupo, con otro individuo que tenga muchas características en común; en algunos casos pueden ser hermanos gemelos, o simplemente individuos de la misma edad, sexo, con condiciones ambientales semejantes, etc. También se usa este diseño cuando se comparan dos métodos de medición.

Se utiliza el apareamiento para controlar fuentes de variación ajenas al experimento, que podrían influir en los resultados del mismo.

En este caso los datos no se presentan como dos muestras independientes, sino como una muestra de pares de v.a.:

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n),$$

Se calculan las diferencias:

$$D_1 = X_1 - Y_1, D_2 = X_2 - Y_2, \ldots, D_n = X_n - Y_n,$$

y vamos a suponer que estas diferencias tienen distribución normal, es decir, $D_i \sim N(\mu_D, \sigma_D^2)$, donde $\mu_D = \mu_1 - \mu_2$.

Entonces, para construir un intervalo de confianza para μ_D se utiliza la función pivote:

$$T = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}}$$

que tiene distribución Student con $n - 1$ grados de libertad, donde $\overline{D} = \frac{1}{n} \sum_{i=1}^{n} D_i$ y $S_D = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (D_i - \overline{D})^2}$. El intervalo de confianza de nivel $1 - \alpha$ para $\mu_D = \mu_1 - \mu_2$ resulta:

$$IC_{1-\alpha}(\mu_1 - \mu_2) = \left(\overline{D} - t_{\alpha/2,n-1} \times \frac{S_D}{\sqrt{n}}, \overline{D} + t_{\alpha/2,n-1} \times \frac{S_D}{\sqrt{n}} \right)$$

Ejemplo 7.7

Se quiere comparar dos métodos de laboratorio. La concentración de plomo ($\mu g/l$) de cada una de cinco muestras es determinada por dos métodos diferentes, con los resultados que se muestran en la siguiente tabla:

muestra	1	2	3	4	5
oxidación húmeda	71	61	50	60	52
extracción directa	76	68	48	57	61

Suponiendo normalidad para la diferencia entre los métodos de oxidación húmeda y de extracción directa, calculemos el intervalo de confianza de nivel 0.995.

Sea la m.a. $D_1, D_2, \ldots, D_5$ con cada D_i que indica la diferencia de concentración de plomo ($\mu g/l$) entre los métodos ("oxidación húmeda-extracción directa") de la muestra i, cuya distribución es $N(\mu_D, \sigma_D^2)$. Los valores muestrales de las v.a. D_i son: -5, -7, 2, 3 y -9, de los que resulta $\overline{d} = -3.2$, $s_d = 5.4037$ y además tenemos por la Tabla que $t_{0.0025,4} = 5.5976$. Entonces

el intervalo de confianza de nivel 0.995 será:

$$IC_{0.995}(\mu_1 - \mu_2) = \left(-3.2 - 5.5976 \times \frac{5.4037}{\sqrt{5}}, -3.2 + 5.5976 \times \frac{5.4037}{\sqrt{5}}\right) = (-16.2586, 9.8586)$$

■

Para realizar un test el estadístico de prueba para la hipótesis $H_0 : \mu_D = \Delta_0$ es:

$$T = \frac{\overline{D} - \Delta_0}{S_D/\sqrt{n}}$$

que, si H_0 es verdadera, tiene distribución Student con $n-1$ grados de libertad. Resumimos a continuación los casos de tests para muestras apareadas:

RESUMEN 7.4

Para la muestra apareda $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ con distribución normal conjunta, definiendo $D_i = X_i - Y_i$, obtenemos $D_i \sim N(\mu_D, \sigma_D^2)$.			
Hipótesis nula	$H_0 : \mu_D = \Delta_0$		
Valor del estadístico de prueba	$t = \frac{\overline{d} - \Delta_0}{s_D/\sqrt{n}}$		
Hipótesis alternativa	$H_A : \mu_D < \Delta_0$	$H_A : \mu_D > \Delta_0$	$H_A : \mu_D \neq \Delta_0$
Región de rechazo con nivel α	$t < -t_{\alpha,n-1}$	$t > t_{\alpha,n-1}$	$\lvert t\rvert > t_{\alpha/2,n-1}$

Ejemplo 7.8

Se dan los niveles de colesterol (mg/dl) en suero para 12 sujetos, antes y después de un programa combinado de dieta y ejercicio. Se desea medir la efectividad del tratamiento para reducir el colesterol, expresada por la diferencia de valores medios entre “antes” y “después”. Se considera que el tratamiento es efectivo si reduce los valores medios de colesterol en más de 50 mg/dl. Considerar normalidad para las diferencias.

Sujeto	1	2	3	4	5	6	7	8	9	10	11	12
antes	281	285	305	298	356	287	273	287	289	317	324	281
después	210	216	239	238	289	232	227	223	240	237	256	206
Diferencia	71	69	66	60	67	55	46	64	49	80	68	75

Sea la m.a. $D_1, D_2, \ldots, D_{12}$ con cada D_i que indica la diferencia de los niveles de colesterol (mg/dl) en suero antes y después del tratamiento en el sujeto i (las diferencias son “antes-después”) cuya distribución es $N(\mu_D, \sigma_D^2)$. De los valores observados para estas diferencias, ya calculadas en la tabla de arriba, se obtienen $\overline{d} = 64.1667$ y $s_d = 10.116$. Luego, el

problema queda planteado como: $H_0 : \mu_D = 50$ vs. $H_A : \mu_D > 50$, el estadístico de prueba será:

$$T = \frac{\overline{D} - 50}{S_D/\sqrt{n}}$$

que bajo H_0 tiene distribución de Student con $n-1$ grados de libertad y la regla de decisión será: rechazar $H_0 : \mu_D = 50$ a favor de $H_A : \mu_D > 50$ cuando $t > t_{\alpha,11}$.

El valor del estadístico de prueba es $t = 4.8512$, vemos que el valor del estadístico de prueba es mayor que todos los valores críticos que tenemos tabulados para 11 grados de libertad en nuestra Tabla, luego el valor-$p < 0.0005$. Esto significa que hay fuerte evidencia de que el tratamiento es efectivo y reduce los valores medios de colesterol en más de 50 md/dl.

■

EJERCICIO 7.4

La adición de imágenes médicas computarizadas a una base de datos promete proporcionar grandes recursos para médicos. Sin embargo, existen otros métodos de obtener tal información, de modo que el tema de eficiencia de acceso tiene que ser investigado. Un artículo de revista reportó sobre un experimento, en el cual a 10 profesionales médicos expertos en la computadora se les tomó el tiempo mientras recuperaban una imagen de una biblioteca de diapositivas y mientras recuperaban la misma imagen de una base de datos de una computadora con conexión a la Web.

Sujeto	1	2	3	4	5	6	7	8	9	10
Diapositiva	32	35	40	25	20	30	35	52	40	41
Digital	20	16	15	15	10	20	7	16	15	13
Diferencia	12	19	25	10	10	10	28	36	25	28

Calcule el intervalo de confianza del 96 % para la diferencia de medias. Haga los supuestos necesarios.

Intervalo de confianza y test para la diferencia de medias con muestras grandes

Cuando tenemos dos muestras independientes, pero desconocemos la distribución de los datos, si las muestras son "grandes" se puede usar la aproximación del Teorema Central del Límite (TCL) como en el caso de una muestra.

Si tenemos dos muestras independientes $X_1, X_2, \ldots, X_{n_1}$ e $Y_1, Y_2, \ldots, Y_{n_2}$ con distribuciones desconocidas y tamaños muestrales, n_1 y n_2, grandes, con $E(X_i) = \mu_1$ y $E(Y_i) = \mu_2$ y queremos

construir un intervalo de confianza para $\mu_1 - \mu_2$, la función pivote que podemos utilizar es:

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Cuando n_1 y n_2 son "grandes", Z tiene una distribución aproximadamente $N(0, 1)$. Entonces el intervalo de confianza para $\mu_1 - \mu_2$, de nivel aproximado $1 - \alpha$, será:

$$IC_\alpha(\mu_1 - \mu_2) = \left(\overline{X} - \overline{Y} - z_{\alpha/2} \times \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \overline{X} - \overline{Y} + z_{\alpha/2} \times \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right)$$

Ejemplo 7.9

Motivados por el conocimiento de la existencia de una gran cantidad de textos polémicos que sugieren que el estrés, la ansiedad y la depresión son dañinos para el sistema inmunológico, se condujo a un estudio en el que se consideró un grupo de pacientes con HIV positivo y otros con HIV negativo. Los datos fueron registrados con una amplia variedad de mediciones médicas, inmunológicas, psiquiátricas y neurológicas, una de las cuales corresponde al número de células CD4+ en la sangre. El número promedio de células CD4+ para 112 individuos con infección por HIV fue de 401.8 con una desviación estándar de 226.4 y para los 75 individuos sin infección por HIV, la media y la desviación estándar fueron de 828.2 y 274.9, respectivamente. Se pretende elaborar un intervalo de confianza de 99 % para las diferencias de las medias de las poblaciones.

Modelizando tenemos una m.a. $X_1, X_2, \ldots, X_{112}$ con cada X_i que indica el número de células CD4+ en la sangre en el paciente i con HIV positivo siendo $E(X_i) = \mu_1$ y otra m.a $Y_1, Y_2, \ldots, Y_{75}$ donde Y_j indica el número de células CD4+ en la sangre en el paciente j con HIV negativo siendo $E(Y_j) = \mu_2$. Las v.a. X_i e Y_j son independientes. No hay información respecto a la distribución del número de células CD4+ en la sangre (se desconoce la distribución de las v.a. X_i e Y_j). Sin embargo, como el tamaño de las muestras es grande, el Teorema Central del Límite asegura que la distribución muestral de las diferencias entre las medias muestrales siguen una distribución aproximadamente normal.

Entonces tenemos $\overline{x} = 401.8$, $\overline{y} = 828.2$, $s_1 = 226.4$, $s_2 = 274.9$ y $z_{0.005} = 2.5758$. Por lo tanto, el intervalo de confianza de nivel aproximado del 99 % para la diferencia de las medias es:

$$IC_{0.99}(\mu_1 - \mu_2) = 401.8 - 828.2 \pm 2.5758 \times \sqrt{\frac{226.4^2}{112} + \frac{274.9^2}{75}} = (-524.998, -327.802)$$

Se tiene la seguridad de 99 % de que el promedio de células CD4+ en pacientes con HIV negativo supera a la media de los pacientes con HIV positivo en un valor entre 327.802 y 524.998.

■

Si se quiere realizar un test para comparar las medias, donde la hipótesis nula es $H_0 : \mu_1 - \mu_2 = \Delta_0$, el estadístico de prueba que podemos usar es:

$$Z = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

que, cuando H_0 es verdadera y n_1 y n_2 son "grandes", tiene una distribución aproximadamente $N(0,1)$. A continuación, resumimos los tests:

RESUMEN 7.5

Para las m.a. independientes $X_1, X_2, \ldots, X_{n_1}$ e $Y_1, Y_2, \ldots, Y_{n_2}$ con distribuciones desconocidas donde n_1 y n_2 son "grandes".			
Hipótesis nula	$H_0: \mu_1 - \mu_2 = \Delta_0$		
Valor del estadístico de prueba	$z = \dfrac{\overline{x} - \overline{y} - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$		
Hipótesis alternativa	$H_A: \mu_1 - \mu_2 < \Delta_0$	$H_A: \mu_1 - \mu_2 > \Delta_0$	$H_A: \mu_1 - \mu_2 \neq \Delta_0$
Región de rechazo con nivel aproximado α	$z < -z_\alpha$	$z > z_\alpha$	$\lvert z\rvert > z_{\alpha/2}$

Ejemplo 7.10

En un estudio de factores que se consideran responsables de los efectos adversos del tabaquismo sobre la reproducción humana, se midieron los niveles de cadmio (nanogramos por gramo) en el tejido de la placenta. En una muestra de 140 madres que fumaban se obtuvo una media de 16.72 ng y un desvío de 6.19 ng. En otra muestra aleatoria independiente de 180 mujeres no fumadoras se obtuvo una media de 18.41 ng y un desvío de 6.81 ng. ¿Hay suficiente evidencia para afirmar que el tabaco reduce el nivel de cadmio en el tejido de la placenta?

Modelizamos: una m.a. $X_1, X_2, \ldots, X_{140}$ con cada X_i que indica el nivel de cadmio (nanogramos por gramo) en el tejido de la placenta en la madre i fumadora siendo $E(X_1) = \mu_1$ y otra m.a $Y_1, Y_2, \ldots, Y_{180}$ que indica el nivel de cadmio (nanogramos por gramo) en el tejido de la placenta en la madre j no fumadora siendo $E(Y_j) = \mu_2$. Las v.a. X_i e Y_j son independientes entre sí y no hay información con respecto a la distribución del nivel de cadmio (nanogramos por gramo) en el tejido de la placenta, es decir que se desconoce la distribución de las v.a. X_i e Y_j. Sin embargo, como el tamaño de las muestras es grande, el Teorema Central del Límite asegura que la distribución muestral de las diferencias entre las medias muestrales siguen una distribución aproximadamente normal.

El planteo del test que deseamos es:

$$H_0: \mu_1 = \mu_2 \quad vs. \quad H_A: \mu_2 > \mu_1$$

o en forma equivalente:

$$H_0: \mu_1 - \mu_2 = 0 \quad vs. \quad H_A: \mu_1 - \mu_2 < 0$$

Para este problema el estadístico de prueba será:

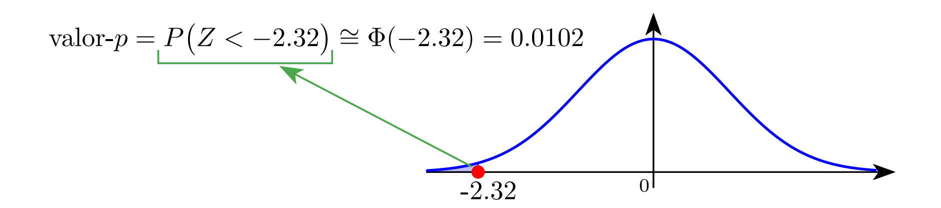
$$Z = \frac{\overline{X} - \overline{Y} - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Cuando $H_0 : \mu_1 - \mu_2 = 0$ es verdadera, la distribución de este estadístico de prueba se aproxima a una normal (por TCL) y la regla de decisión será rechazar H_0 si el valor $z = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} < -z_\alpha$.

El valor del estadístico de prueba es:

$$z = \frac{16.72 - 18.41}{\sqrt{\frac{6.19^2}{140} + \frac{6.81^2}{180}}} = -2.32$$

Calculemos el valor-p para poder responder:



esto significa que, hasta con un nivel 0.0102, podemos afirmar que el tabaco reduce el nivel de cadmio en el tejido de la placenta.

■

Ahora, en cambio, si tenemos muestras apareadas grandes y no conocemos la distribución, también se calculan las diferencias y se trabaja como en el caso de una muestra grande aplicando el TCL. Es decir, la función pivote será

$$Z = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}}$$

que, cuando n es grande, tiene distribución aproximada $N(0, 1)$, entonces el intervalo de confianza para $\mu_D = \mu_1 - \mu_2$ de nivel aproximado $1 - \alpha$ será:

$$IC_{1-\alpha}(\mu_D) = \left(\overline{D} - z_{\alpha/2} \times \frac{S_D}{\sqrt{n}}, \overline{D} + z_{\alpha/2} \times \frac{S_D}{\sqrt{n}}\right)$$

Ejemplo 7.11

Las personas que padecen el síndrome de Reynaud están propensas a sufrir un deterioro repentino de la circulación sanguínea en los dedos de las manos y de los pies. En un experimento para estudiar el grado de este deterioro, cada uno de los 58 sujetos sumergió un dedo índice en agua y se midió la producción de calor resultante ($cal/cm^2/min$). La diferencia de producción de calor promedio antes y después de haber puesto el dedo en el agua fue 0.64 y un desvío estándar de 0.2. Construya el intervalo de confianza del 98 % para la diferencia promedio de calor resultante antes y después de sumergir el dedo en el agua.

Sean D_i las diferencias de producción de calor antes y después de haber puesto manos y pies en el agua, $i = 1, \ldots, 58$ siendo $E(D_i) = \mu_D$. La distribución de las D_i es desconocida pero la muestra es grande y podemos usar TCL.

La función pivote es:

$$Z = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}}$$

que tiene distribución aproximada $N(0,1)$. Tenemos $\overline{d} = 0.64$, $s_d = 0.2$ y $z_{0.01} = 2.3263$ entonces el intervalo es:

$$IC_{0.98}(\mu_D) = \left(0.64 - 2.3263 \times \frac{0.2}{\sqrt{58}}, 0.64 + 2.3263 \times \frac{0.2}{\sqrt{58}}\right) = (0.5789, 0.7011)$$

■

Para realizar un test, el estadístico de prueba es

$$Z = \frac{\overline{D} - \Delta_0}{S_D/\sqrt{n}}$$

que, si $H_0 : \mu_D = \Delta_0$ es verdadera y n es grande, tiene distribución aproximada $N(0,1)$. Resumimos a continuación los distintos casos de tests:

RESUMEN 7.6

Para la muestra apareada $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, con distribución desconocida y n grande, definiendo $D_i = X_i - Y_i$, donde $E(D_i) = \mu_D$.			
Hipótesis nula	$H_0 : \mu_D = \Delta_0$		
Valor del estadístico de prueba	$z = \dfrac{\overline{d} - \Delta_0}{s_D/\sqrt{n}}$		
Hipótesis alternativa	$H_A : \mu_D < \Delta_0$	$H_A : \mu_D > \Delta_0$	$H_A : \mu_D \neq \Delta_0$
Región de rechazo con nivel aproximado α	$z < -z_\alpha$	$z > z_\alpha$	$\lvert z \rvert > z_{\alpha/2}$

Ejemplo 7.12

Se requiere información sobre la postura de la mano y las fuerzas generadas por los dedos durante la manipulación de varios objetos cotidianos para diseñar prótesis de alta tecnología para la mano. Cierto artículo reportó que para una muestra de 110 mujeres la diferencia entre la fuerza de opresión con cuatro dedos (N) entre la mano derecha e izquierda fue de 98.1 N y la desviación estándar de 143.2 N. ¿Existe evidencia suficiente para concluir que hay diferencia entre la fuerza promedio verdadera en cada mano?

Sean D_i las diferencias entre la fuerza de opresión con cuatro dedos (N) entre la mano derecha e izquierda, $i = 1, \ldots, 110$, siendo $E(D_i) = \mu_D$. La distribución de las D_i es desconocida pero la muestra es grande y podemos usar TCL. Las hipótesis son:

$$H_0 : \mu_D = 0 \quad vs. \quad H_A : \mu_D \neq 0$$

Para este problema el estadístico de prueba será:

$$Z = \frac{\overline{D} - 0}{S_D/\sqrt{110}}$$

con distribución aproximada $N(0,1)$, si H_0 es verdadera.

Reemplazando con los datos, el valor del estadístico de prueba es $z = \frac{98.1}{143.2/\sqrt{110}} = 7.1849$. Luego valor-$p = P(|Z| > 7.18) \cong 2 \times (1 - \Phi(7.18)) \cong 0$. Es decir, hay mucha evidencia contra H_0, con lo cual se puede afirmar que hay diferencia entre la fuerza promedio verdadera en cada mano.

■

EJERCICIO 7.5

Un programa de pérdida de peso afirma que después de tres meses de tratamiento los pacientes pierden en media 4 kg. Se estudió una muestra de 60 participante de ese programa y se observó que la media de las diferencias de pesos fue de 4.6 kg y su desvió fue de 1.2 kg. ¿Estos datos confirman dicha afirmación?

Inferencias en relación con la diferencia entre proporciones (muestras grandes)

Ya vimos al tratar el tema intervalo de confianza y test para una proporción que, si tenemos una v.a. X con distribución $B(n,p)$ y n es suficientemente grande, la distribución de:

$$Z = \frac{\widehat{p} - p}{\sqrt{p\,(1-p)/n}}$$

es aproximadamente $N(0,1)$ por TCL. Ahora, si tenemos dos v.a. X e Y independientes y con distribuciones $B(n_1,p_1)$ y $B(n_2,p_2)$ respectivamente, donde n_1 y n_2 son grandes, definimos $\widehat{p}_1 = X/n_1$ y $\widehat{p}_2 = Y/n_2$, el estimador obvio para $p_1 - p_2$ es $\widehat{p}_1 - \widehat{p}_2$, y se puede ver que:

$$E(\widehat{p}_1 - \widehat{p}_2) = p_1 - p_2 \quad \text{y} \quad Var(\widehat{p}_1 - \widehat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

entonces:

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2 - \big(p_1 - p_2\big)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

tiene distribución aproximadamente $N(0,1)$. Si se desea construir un intervalo de confianza para $p_1 - p_2$, no podemos usar Z como función pivote, pero podemos modificarla y usar:

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2 - \big(p_1 - p_2\big)}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2\,(1-\widehat{p}_2)}{n_2}}}$$

que también tiene distribución aproximadamente $N(0,1)$.

Entonces el intervalo de confianza de nivel aproximado $1-\alpha$ resulta:

$$IC_{1-\alpha}(p_1 - p_2) = \widehat{p}_1 - \widehat{p}_2 \pm z_{\alpha/2} \times \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2\,(1-\widehat{p}_2)}{n_2}}$$

Ejemplo 7.13

Se encuestaron 67 médicos y 133 enfermeras con familiares fármaco-dependientes. El propósito del estudio era evaluar la influencia en los médicos y enfermeras de estar estrechamente involucrados con una o más personas fármaco-dependientes. Se obtuvo que 52 médicos y 89 enfermeras dijeron que vivir con personas fármaco-dependientes afectaban adversamente su trabajo. Deseamos construir un intervalo de confianza de 98 % para la diferencia entre las proporciones de médicos y enfermeras que estén adversamente afectados por vivir con personas fármaco-dependientes.

Modelizamos el problema: sean la v.a. X el número de médicos, entre los 67 que se encuestaron, que son afectados en su trabajo, $X \sim B(67, p_1)$ y la v.a. Y el número de enfermeras, entre las 133 que se encuestaron, que son afectadas en su trabajo, $Y \sim B(133, p_2)$. Entonces tenemos: $x = 52$, $y = 89$, resultando $\widehat{p}_1 = 52/67$ y $\widehat{p}_2 = 89/133$. Y además $z_{0.01} = 2.3263$. Entonces el intervalo de confianza de nivel aproximado 0.98 resulta:

$$\begin{aligned} IC_{0.98}(p_1 - p_2) &= \frac{52}{67} - \frac{89}{133} \pm 2.3263 \times \sqrt{\frac{52/67\;(1-52/67)}{67} + \frac{89/133\;(1-89/133)}{133}} \\ &= (-0.0449, 0.2587) \end{aligned}$$

■

Ahora, si se desea hacer un test de hipótesis donde $H_0 : p_1 = p_2$, para definir la función pivote tendremos en cuenta que, bajo H_0, las dos proporciones son iguales y podemos llamar p a ese valor (es decir, $p = p_1 = p_2$). Entonces:

$$E(\widehat{p}_1 - \widehat{p}_2) = p_1 - p_2 = 0 \quad \text{y} \quad Var(\widehat{p}_1 - \widehat{p}_2) = \frac{p\,(1-p)}{n_1} + \frac{p\,(1-p)}{n_2} = p\,(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Luego, si H_0 es verdadera, el estadístico:

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{p\,(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

tiene distribución aproximadamente $N(0,1)$ y puede usarse como estadístico de prueba. Pero como no conocemos el valor p lo reemplazamos por $\widehat{p} = (X_1 + X_2)/(n_1 + n_2)$, que también tiene distribución aproximadamente $N(0,1)$. Resumimos a continuación los distintos casos de tests:

RESUMEN 7.7

Para las v.a. X e Y independientes, donde $X \sim B(n_1, p_1)$ e $Y \sim B(n_2, p_2)$, con n_1 y n_2 grandes.			
Hipótesis nula	$H_0 : p_1 = p_2$		
Valor del estadístico de prueba	$z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}\,(1-\widehat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$		
Hipótesis alternativa	$H_A : p_1 < p_2$	$H_A : p_1 > p_2$	$H_A : p_1 \neq p_2$
Región de rechazo con nivel aproximado α	$z < -z_\alpha$	$z > z_\alpha$	$\lvert z \rvert > z_{\alpha/2}$

Ejemplo 7.14

Se realizó un estudio con 2720 pacientes que habían sufrido un infarto cardíaco. El estudio asignó aleatoriamente cada individuo a un grupo de tratamiento, que recibieron una dosis diaria de aspirina y un grupo control, tratado con un placebo. Se realizó un seguimiento durante 3 años. La tabla muestra los resultados obtenidos:

Grupo	Número de muertos	Tamaño de la muestra
Placebo	56	1368
Aspirina	36	1352

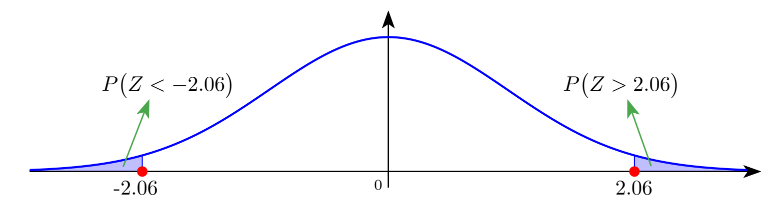
El objetivo de este estudio era determinar si la toma diaria de aspirina en pacientes que ya habían sufrido un ataque cardíaco, modifica la tasa de mortalidad para futuros infartos.

Podemos definir X como el número de muertes por infarto durante el período del estudio en pacientes que recibieron placebo, $X \sim B(1368, p_1)$ e Y como el número de muertes por infarto durante el período del estudio en pacientes que recibieron aspirina, $Y \sim B(1352, p_2)$. Se quiere testear: $H_0 : p_1 = p_2$ vs.$H_A : p_1 \neq p_2$. Si H_0 es verdadera, el estadístico:

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}\,(1-\widehat{p})\left(\frac{1}{1368}+\frac{1}{1352}\right)}}$$

tiene distribución aproximadamente $N(0, 1)$. Aquí, $\widehat{p}_1 = 0.0409$, $\widehat{p}_2 = 0.0266$ y $\widehat{p} = 0.0338$. El valor del estadístico de prueba es $z = 2.06$ y el valor-p para esta prueba bilateral es:

$$\text{valor-}p = P(|Z| > 2.06) = P\big(Z > 2.06\big) + P\big(Z < -2.06\big) \cong 0.0394$$



Esto significa que hay evidencia para afirmar que hay diferencias entre las tasas de mortalidad.

EJERCICIO 7.6

Como parte de un estudio para identificar factores de riesgo coronario entre hombres y mujeres en una clínica, se estudió la proporción de diabéticos en hombres y mujeres. De las 215 mujeres 58 tenían diabetes al igual que 21 de los 114 hombres.

1. Elabore un intervalo de confianza del 99 % para la diferencia de proporciones de diabetes en hombres y mujeres.

2. Se pretende saber si es posible concluir, con base a estos datos, que en general, la prevalencia de diabetes es superior en mujeres con nivel $\alpha = 0.05$. Aproxime el valor-p.

Relación entre intervalo de confianza y test de hipótesis

Sea $X_1, X_2, \ldots, X_n$ una m.a. de una distribución $F(\theta)$ y sea $IC_{1-\alpha}(\theta)$ un intervalo de confianza de nivel $1-\alpha$ para θ, esto significa que:

$$P\Big(\theta \in IC_{1-\alpha}(\theta)\Big) = 1-\alpha,$$

como vimos en el Capítulo 5.

Consideremos el problema de test de hipótesis:

$$H_0 : \theta = \theta_0 \quad vs. \quad H_A : \theta \neq \theta_0$$

si H_0 es verdadera, entonces:

$$P\Big(\theta_0 \in IC_{1-\alpha}(\theta)\Big) = 1-\alpha$$

lo cual implica que:

$$P\Big(\theta_0 \notin IC_{1-\alpha}(\theta)\Big) = \alpha$$

Entonces, podemos establecer la siguiente regla de decisión:

$$\text{rechazar } H_0 : \theta = \theta_0 \text{ a favor de } H_A : \theta \neq \theta_0 \text{ cuando } \theta_0 \notin IC_{1-\alpha}(\theta),$$

de este modo construimos un test bilateral de nivel α.

Ejemplo 7.15

Consideremos los datos del Ejemplo 5.13, allí construimos un intervalo de confianza para la concentración de ion nitrato en una muestra de agua.

Si estamos interesados en saber si la verdadera concentración es 53 $\mu g/ml$, las hipótesis a

contrastar son:

$$H_0 : \mu = 53 \quad vs. \quad H_A : \mu \neq 53$$

podemos construir un test a partir del intervalo calculado antes. La regla de decisión será rechazar H_0 si $53 \notin IC_{1-\alpha}(\theta)$. El intervalo de 95 % de confianza obtenido fue (48.98, 52.16), y como el valor 53 no está dentro de ese intervalo, podemos rechazar H_0 con nivel $\alpha = 0.05$ y la conclusión, a ese nivel, será que la concentración de ion nitrato en esa muestra no es 53 $\mu g/ml$.

■

Referencias

Agresti, A. & Franklin, C. A. (2009). *Statistics: The Art and Science of learning from Data.* Pearson New International edition.

Altman, D. G. (1990). *Practical Statistics for Medical Research.* Published by Chapman & Hall.

Daniel, W. (2002). *Bioestadística: Base para el análisis de las ciencias de la salud.* Ed. Limusa Wiley.

Devore Jay, L. (2001). *Probabilidad y Estadística para Ingeniería y Ciencias.* Ed. Books/Cole Publishing Company.

Dixon, W. & Massey, F. (1970). *Introducción al Análisis Estadístico.* México. Libros Mc Graw-Hill.

Maronna, R. (1995). *Probabilidad y Estadística Elementales para Estudiantes de Ciencias.* Buenos Aires.Ed. Exactas.

Mendenhall, W., Beaver, R. J. & Beaver, B. M. (2006). *Introducción a la Probabilidad y Estadística.* México. Cengage Learning Editores.

Ross, S. M. (1987). *Introduction to Probability and Statistics for Enginees and Scientists.* Published by John Wiley & Sons.

Wackerly, D. D., Mendenhall, W. & Scheaffer, R. L. (2010). *Estadística Matemática con aplicaciones.* México. Cengage Learning Editores.

Walpole, R. E. & Myers, R. H. (2007). *Probabilidad y Estadística para Ingeniería y Ciencias.* México. Ediciones McGraw-Hill.

CAPÍTULO 8

Modelo de regresión lineal

Introducción

En capítulos anteriores, aún cuando analizamos muestras de dos distribuciones, nos interesamos en comparar sus parámetros, pero no usamos la información de una muestra para hacer inferencias sobre la otra distribución. En este capítulo, veremos cómo obtener información sobre una variable (variable respuesta) a partir del conocimiento de otras (variables explicativas).

En cursos de Análisis Matemático se han estudiado relaciones determinísticas, si dos variables x e y están relacionadas de esa manera, sabiendo el valor de x, podemos conocer exactamente el valor de y.

En muchas aplicaciones encontramos variables que parecen estar relacionadas, aunque no de manera determinística. Esto significa que, saber el valor de las variables explicativas no nos permite conocer exactamente el valor de la variable respuesta, ya que ésta puede considerarse una variable aleatoria; pero se pueden hacer inferencias sobre la distribución de probabilidad de dicha variable. Por ejemplo, sean x la edad de un niño e y su talla, sabemos que la talla depende de la edad, pero también depende de muchas otras condiciones; para una determinada edad, la talla puede considerarse como una v.a.

Los modelos de regresión lineal, estudian la relación entre dos o más variables relacionadas en una forma no determinística.

El caso más simple que estudiaremos en este capítulo, es el modelo de regresión lineal simple, donde tenemos una única variable explicativa.

Modelo de regresión lineal simple

La relación matemática determinística más simple entre dos variables x e y, es una relación lineal $y = \beta_0 + \beta_1 x$. El conjunto de pares (x, y) que verifican esta relación, determinan una recta con pendiente β_1 que corta al eje y en β_0.

En esta sección vamos a estudiar una relación lineal no determinística entre dos variables. La variable cuyo valor fija el experimentador será denotada por x y se llamará variable independiente, pronosticadora o variable explicativa. Con x fija, la segunda variable es aleatoria y se la denomina variable dependiente o respuesta; esta variable aleatoria la designamos como Y y a su valor observado lo designamos con y.

Comenzaremos analizando el siguiente ejemplo:

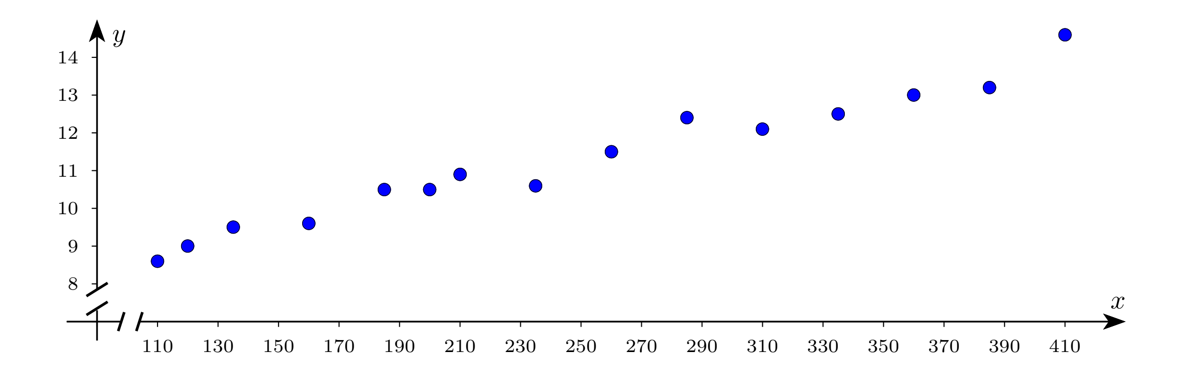
Ejemplo 8.1

Los siguientes valores provienen de un estudio sobre calidad del aire en una ciudad, son las lecturas sobre el volumen de tránsito (en número de automóviles por hora) y la concentración de monóxido de carbono, en un punto de muestreo.

Vol	100	110	125	150	175	190	200	225
CO	8.6	9.0	9.5	9.6	10.5	10.5	10.9	10.6

Vol	250	275	300	325	350	375	400
CO	11.5	12.4	12.1	12.5	13.0	13.2	14.6

Un primer paso en el análisis de regresión, es construir una gráfica de puntos de los datos observados. En una gráfica como esta, cada (x_i, y_i) está representado como un punto en un sistema de coordenadas bidimensional.



Vemos que los puntos parecen estar bastante próximos a una recta, y podemos aceptar que la relación entre las variables es *aproximadamente lineal*. Esto significa que la concentración de monóxido de carbono (y_i) parece estar relacionada linealmente con el volumen de tránsito (x_i).

En un problema general se realizan varias observaciones. Sean $x_1, x_2, \ldots, x_n$ los valores de

la variable independiente y sean $y_1, y_2, \ldots, y_n$ respectivamente, los valores observados de la variable dependiente asociados con los x_i. Los datos disponibles se componen entonces de los n pares $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, que analizaremos suponiendo que se verifican las hipótesis del siguiente modelo:

Definición:

Para un conjunto de observaciones $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, donde y_i son valores observados de variables aleatorias Y_i relacionadas con las x_i, el **modelo de regresión lineal simple** se define como:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{8.1}$$

donde β_0 y β_1 son parámetros fijos y los ϵ_i son variables aleatorias que cumplen:

- $E(\epsilon_i) = 0$, $i = 1, \ldots, n$
- $var(\epsilon_i) = \sigma^2$, $i = 1, \ldots, n$
- ϵ_i y ϵ_j son independientes entre sí, para $i \neq j$.

Esto significa que, para cada valor de la variable *independiente o explicativa* x_i, la variable *dependiente* o variable *respuesta* Y_i, es una variable aleatoria tal que:

- $E(Y_i) = \beta_0 + \beta_1 x_i$, $i = 1, \ldots, n$
- $var(Y_i) = \sigma^2$, $i = 1, \ldots, n$
- Y_i e Y_j son independientes entre sí, para $i \neq j$.

Estimación de los parámetros $\boldsymbol{\beta_0}$ y $\boldsymbol{\beta_1}$

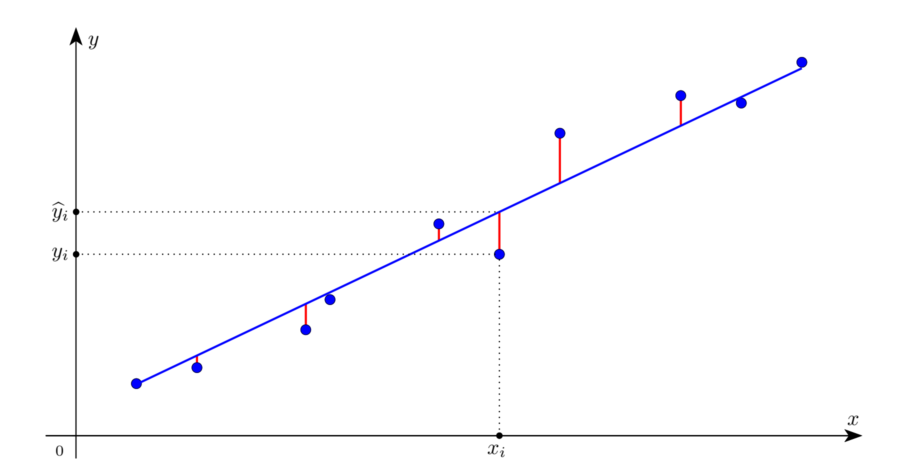
Según el modelo, los puntos observados se distribuyen aleatoriamente alrededor de la verdadera recta $y = \beta_0 + \beta_1 x$. Para estimar dicha recta se utilizará el método de **mínimos cuadrados**.

Consideremos una posible recta estimada $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$, para cada valor x_i sea $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ el correspondiente valor sobre la recta estimada.

Definiendo los **residuos** como:

$$r_i = y_i - \widehat{y}_i = y_i - \big(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\big), \tag{8.2}$$

muestran la discrepancia entre los valores observados y_i y los correspondientes valores estimados $\widehat{y}_i$ para cada x_i, como se muestra en el siguiente gráfico:



Parece lógico pretender que esos residuos sean mínimos. No podemos minimizar cada residuo por separado, pero podemos minimizarlos de manera global, minimizando la suma de los residuos al cuadrado (S_{rr}).

$$S_{rr} = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \left(y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i \right) \right)^2 = h(\widehat{\beta}_0, \widehat{\beta}_1)$$

Entonces, el método consiste en hallar la recta que minimice dicha suma, ahora bien, determinar la recta equivale a determinar su pendiente $\widehat{\beta}_1$ y su ordenada al origen $\widehat{\beta}_0$. Es decir, debemos hallar los valores $\widehat{\beta}_0$ y $\widehat{\beta}_1$ que hagan mínima $S_{rr} = h(\widehat{\beta}_0, \widehat{\beta}_1)$.

Calculando las derivadas parciales de $h(\widehat{\beta}_0, \widehat{\beta}_1)$ respecto de $\widehat{\beta}_0$ y de $\widehat{\beta}_1$ e igualando ambas a cero, se obtiene un sistema de dos ecuaciones, al resolverlo, se llega a la siguiente solución:

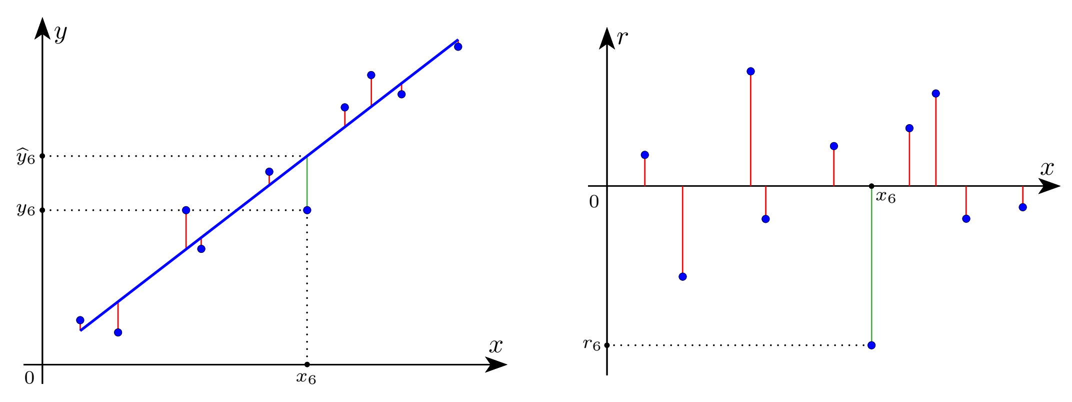
$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{y} \quad \widehat{\beta}_0 = \overline{y} - \overline{x}\widehat{\beta}_1$$

donde $\overline{x} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i$, $\overline{y} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} y_i$, $S_{xx} = \displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2$ y $S_{xy} = \displaystyle\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$.

Definición:

La recta obtenida $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$, se llama **recta de regresión estimada de y en x.**

Una vez estimada la recta de regresión se pueden calcular los residuos (8.2) y hacer un gráfico con los residuos en el eje y. En la siguiente figura se muestran los gráficos de la recta estimada y de los residuos correspondientes.



Este tipo de gráfico nos permite verificar si las hipótesis del modelo son válidas. Si el modelo de regresión es correcto, la gráfica de residuos no debería presentar ningún patrón distinguible.

Estimación de σ^2

La varianza σ^2 mide la variabilidad inherente al modelo de regresión. Una estimación de σ^2 se basará en cuánto se desvían las observaciones de la recta estimada. Entonces, la varianza σ^2 se estima con s_r^2 definido como:

$$s_r^2 = \frac{\sum_{i=1}^{n}\left(y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right)^2}{n-2} = \frac{S_{rr}}{n-2}$$

Una manera simple de calcular S_{rr} es la siguiente:

$$S_{rr} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

donde

$$S_{yy} = \sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$$

La demostración es:

$$\begin{aligned}
S_{rr} &= \sum_{i=1}^{n}\left(y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\right)^2 = \sum_{i=1}^{n}\left[y_i - \left(\overline{y} - \widehat{\beta}_1 \overline{x} + \widehat{\beta}_1 x_i\right)\right]^2 \\
&= \sum_{i=1}^{n}\left[(y_i - \overline{y}) - \left(\widehat{\beta}_1\left(x_i - \overline{x}\right)\right)\right]^2 = \sum_{i=1}^{n}\left[(y_i - \overline{y})^2 - 2\left(y_i - \overline{y}\right)\widehat{\beta}_1\left(x_i - \overline{x}\right) + \widehat{\beta}_1^2\left(x_i - \overline{x}\right)^2\right] \\
&= S_{yy} - 2\widehat{\beta}_1 S_{xy} + \widehat{\beta}_1^2 S_{xx} = S_{yy} - 2\frac{S_{xy}}{S_{xx}}S_{xy} + \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}
\end{aligned}$$

El coeficiente de determinación

Una medida de la variablidad total de las observaciones y_i es la expresión que ya vimos S_{yy}, parte de esa variabilidad es explicada por el modelo de regresión. Pero como el modelo no ajusta perfectamente a los datos, queda una parte de esa variabilidad que es aleatoria y no puede ser explicada por el modelo. La suma de cuadrados de los residuos (S_{rr}), puede considerarse como una medida de esa variabilidad residual que no es explicada por el modelo, obviamente $S_{yy} \geq S_{rr}$. Entonces, el cociente Srr/Syy da la proporción de la variación total en las y_i que no es explicada por el modelo de regresión y $0 \leq Srr/Syy \leq 1$.

Definición:

En un modelo de regresión lineal, se define el **coeficiente de determinación** como:

$$r^2 = 1 - \frac{S_{rr}}{S_{yy}}$$

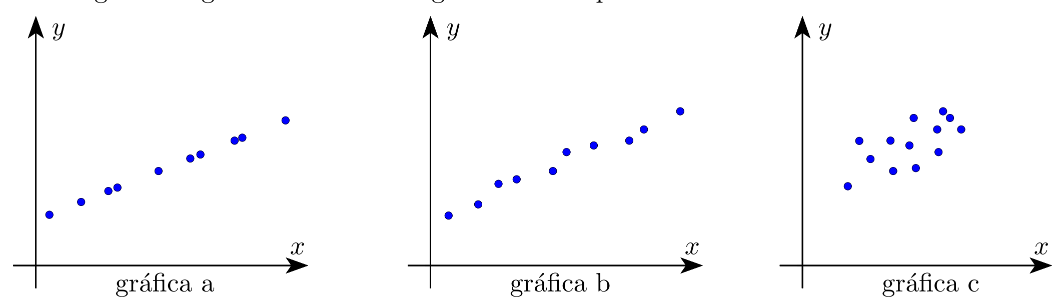
y puede interpretarse como la proporción de la variabilidad total de las y_i que es explicada por el modelo.



Observación:

Es evidente que se cumple: $0 \leq r^2 \leq 1$ y el coeficiente de determinación es una medida de la bondad del ajuste del modelo, un valor de $r^2 = 1$ indicaría un ajuste perfecto.

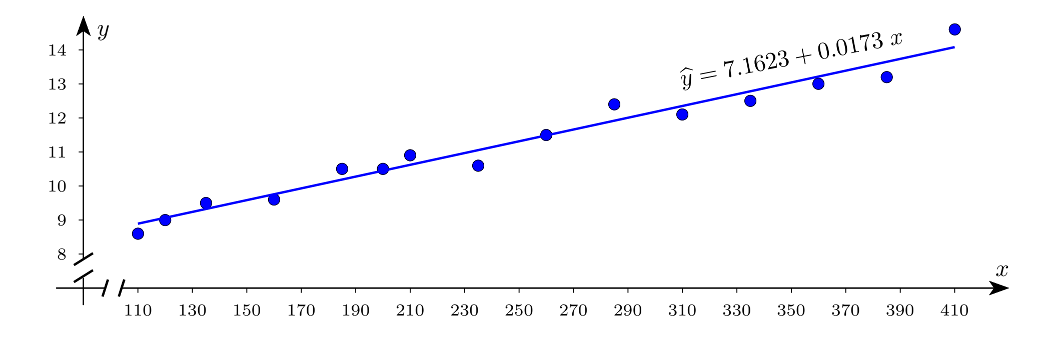
En la siguiente figura se muestran 3 gráficos de dispersión diferentes:



- si todos los puntos caen en la recta de regresión (gráfica a), el $S_{rr} = 0$ y $r^2 = 1$
- si los puntos no caen exactamente en una recta, pero hay poca variabilidad alrededor de ella, el r^2 será próximo 1 (gráfica b)
- si los puntos están más dispersos en relación a la recta de mínimos cuadrados, la variación explicada por la regresión es menor (gráfica c), r^2 no es tan próximo a 1.

Ejemplo 8.2

Con los datos del Ejemplo 8.1, tenemos que: $n = 15$, $\overline{x} = 236.667$, $\overline{y} = 11.233$, $S_{xx} = 134283.333$, $S_{xy} = 2314.167$ y $S_{yy} = 41.293$. Luego, obtenemos $\widehat{\beta}_1 = \frac{2314.167}{134283.333} = 0.0173$ y $\widehat{\beta}_0 = 11.233 - 0.0172 \times 236.667 = 7.1623$, así la recta estimada será:



Recordemos que, la pendiente de una recta es el cambio en la variable y que corresponde a un incremento unitario en la variable x. En este modelo, significaría el cambio esperado en Y (concentración de monóxido de carbono) cuando el volumen de tránsito aumenta en un automóvil por hora.

Estimamos la varianza y el coeficiente de determinación:

$$S_{rr} = 41.293 - \frac{2314.167^2}{134283.333} = 1.1419, \quad s_r^2 = \frac{1.4119}{13} = 0.1086 \quad \text{y} \quad r^2 = 1 - \frac{1.4119}{41.293} = 0.9658$$

Esto significa que el 96.58 % de la variabilidad de la concentración de monóxido de carbono, puede explicarse por el modelo lineal que la relaciona con el volumen de tránsito.

■

EJERCICIO 8.1

1. Demostrar las siguientes igualdades:
   - $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$
   - $S_{xy} = \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}$
   - $S_{yy} = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2$
2. Llegar a las soluciones de mínimos cuadrados.
3. Se realizó un experimento para observar el efecto de la temperatura de almacenamiento en la potencia de un antibiótico. Tres porciones de una onza del antibiótico se almacenaron durante tiempos iguales a cada una de las siguientes temperaturas Fahrenheit: 30°, 50°, 70° y 90°. Las lecturas de potencia observadas al final del período experimental fueron como se muestra en la tabla siguiente.

Temperatura (x)	30	30	30	50	50	50	70	70	70	90	90	90
Lecturas de potencia (y)	38	43	34	32	26	33	19	27	23	14	19	21

El resumen de datos es: $\overline{x} = 60$, $\overline{y} = 27.4167$, $s_{xx} = 6000$, $s_{yy} = 834.9167$ y $s_{xy} = -2050$.

   a. Graficar los puntos, estimar y graficar la recta de regresión.
   b. Calcular la proporción de la variabildad en las lecturas de potencia observadas que puede ser explicada por el modelo de regresión lineal.

Definición y propiedades de los estimadores de los parámetros

En la sección anterior utilizamos el método de mínimos cuadrados para encontrar *estimaciones* de los parámetros β_0 y β_1, que son funciones de las observaciones x_i e y_i. Los valores x_i se supone que son elegidos antes de realizar el experimento y no son aleatorios, en cambio, las y_i son valores observados de las respectivas variables aleatorias Y_i.

Reemplazando en las ecuaciones anteriores los valores y_i por las variables aleatorias Y_i, se obtienen *estimadores* de β_0 y β_1.

Definición:

Los *estimadores de mínimos cuadrados* de los parámetros β_0 y β_1 están dados por:

$$\begin{aligned}\widehat{\beta}_1 &= \frac{S_{xY}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})\left(Y_i-\overline{Y}\right)}{\sum_{i=1}^{n}(x_i-\overline{x})^2},\\ \widehat{\beta}_0 &= \overline{Y}-\overline{x}\widehat{\beta}_1\end{aligned}$$

Del mismo modo, el *estimador* de la varianza se obtiene reemplazando las y_i por las Y_i:

$$\widehat{\sigma}^2 = S_r^2 = \frac{\sum_{i=1}^{n}\left(Y_i-\left(\widehat{\beta}_0+\widehat{\beta}_1 x_i\right)\right)^2}{n-2} = \frac{1}{n-2}\left(S_{YY}-\frac{S_{xY}^2}{S_{xx}}\right)$$

Se puede probar, bajo las suposiciones del modelo (8.1), que los estimadores de mínimos cuadrados $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son insesgados, esto quiere decir que:

$$E(\widehat{\beta}_0)=\beta_0 \quad \text{y} \quad E(\widehat{\beta}_1)=\beta_1 \tag{8.3}$$

y también puede probarse que:

$$var(\widehat{\beta}_0)=\sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right) \quad \text{y} \quad var(\widehat{\beta}_1)=\frac{\sigma^2}{S_{xx}} \tag{8.4}$$

EJERCICIO 8.2

Demostrar (8.3) y (8.4).

Inferencias cuando ϵ_i tiene distribución normal

En muchas aplicaciones se puede suponer que los términos aleatorios ϵ_i tienen distribución Normal, $\epsilon_i \sim N(0,\sigma^2)$, y en consecuencia, las variables aleatorias Y_i tienen distribución Normal, $Y_i \sim N(\beta_0+\beta_1 x_i,\sigma^2)$.

En ese caso, se pueden construir intervalos de confianza y tests de hipótesis para los parámetros.

Intervalo de confianza y test de hipótesis para la pendiente

Como $\sum_{i=1}^{n}(x_i-\overline{x})=0$, podemos ver que:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})\,Y_i}{\sum_{j=1}^{n}(x_j-\overline{x})^2} = \sum_{i=1}^{n}\left(\frac{(x_i-\overline{x})}{\sum_{j=1}^{n}(x_j-\overline{x})^2}\right)Y_i = \sum_{i=1}^{n} c_i\,Y_i$$

es decir, $\widehat{\beta}_1$ es combinación lineal de las Y_i, que son normales e independientes, y en consecuencia, $\widehat{\beta}_1$ tiene distribución normal, es decir:

$$\widehat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx}) \quad \text{entonces} \quad \frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$$

Recordemos que, cuando construíamos un intervalo de confianza para la media de una distribución normal con varianza desconocida, la función pivote utilizada al reemplazar σ por su estimador, tenía distribución de Student con $n-1$ grados de libertad. También se puede demostrar, aunque esta fuera del alcance de este curso, que:

$$\frac{\widehat{\beta}_1 - \beta_1}{s_r/\sqrt{S_{xx}}} \sim T_{(n-2)}$$

Ésta será la función pivote que usaremos para construir un intervalo de confianza para la pendiente, con el mismo procedimiento que ya usamos anteriormente.

Los valores críticos que usamos son $-t_{\alpha/2,n-2}$ y $t_{\alpha/2,n-2}$ y planteamos:

$$P\left(-t_{\alpha/2,n-2} \leq \frac{\widehat{\beta}_1 - \beta_1}{s_r/\sqrt{S_{xx}}} \leq t_{\alpha/2,n-2}\right) = 1 - \alpha$$

y finalmente se llega al intervalo:

$$IC_{1-\alpha}(\beta_1) = \left(\widehat{\beta}_1 - t_{\alpha/2,n-2} \times \frac{s_r}{\sqrt{S_{xx}}}, \widehat{\beta}_1 + t_{\alpha/2,n-2} \times \frac{s_r}{\sqrt{S_{xx}}}\right)$$



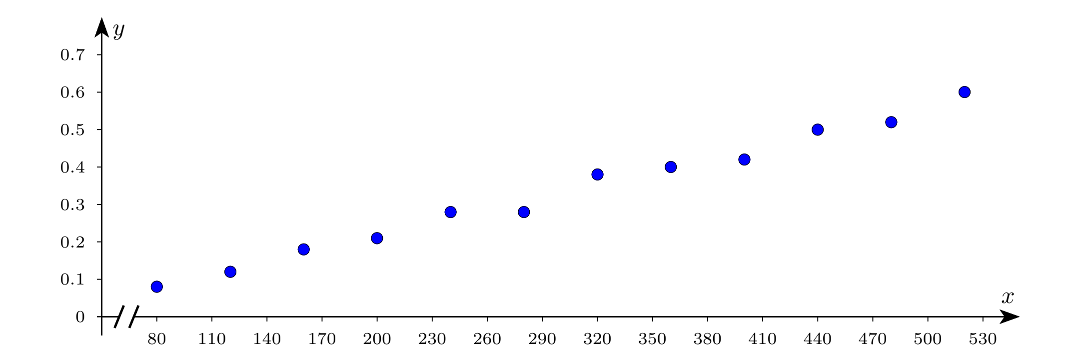
Observación:

La longitud del intervalo para β_1 es $2t_{\alpha/2,n-2}s_r/\sqrt{S_{xx}}$, de modo que la precisión de la estimación para β_1 puede mejorarse eligiendo los valores x_i más dispersos para que S_{xx} sea más grande.

Ejemplo 8.3

Consideremos los siguientes datos que muestran la densidad óptica de cierta substancia (y) a diferentes niveles de concentración (x).

x	80	120	160	200	240	280	320	360	400	440	480	520
y	0.08	0.12	0.18	0.21	0.28	0.28	0.38	0.40	0.42	0.50	0.52	0.60



Vemos que los puntos parecen estar bastante próximos a una recta y podemos aceptar que la relación entre las variables es "aproximadamente lineal". Podemos pensar que para cada concentración (x_i), el valor de la densidad óptica (y_i) es función lineal de x_i más un término aleatorio, que suponemos que tiene distribución normal y se cumplen las hipótesis enunciadas del modelo de regresión lineal.

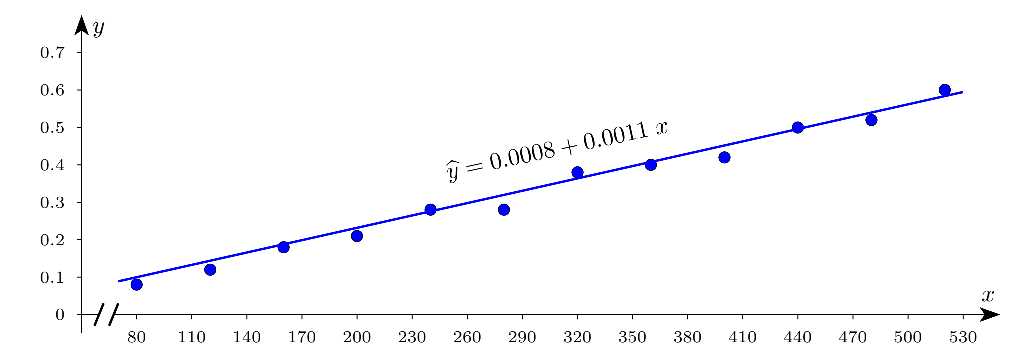
En este caso $n = 12$ y con esos datos: $\overline{x} = 300$, $\overline{y} = 0.3308$, $S_{xx} = 228800$, $S_{yy} = 0.3019$, $S_{xy} = 261.4$, entonces

$$\widehat{\beta}_1 = \frac{261.4}{228800} = 0.0011 \quad \text{y} \quad \widehat{\beta}_0 = 0.3308 - 0.0011 \times 300 = 0.0008$$

de modo que la *recta de regresión estimada* será:

$$\widehat{y} = 0.0008 + 0.0011\, x$$

esto significa que al aumentar la concentración en 1 unidad, la densidad óptica aumenta en 0.0011. Si graficamos estos valores junto con la recta de regresión estimada tenemos:



Para estimar σ según esos datos: $S_{rr} = 0.3019 - 261.4^2/228800 = 0.0032$ y luego, $s_r = \sqrt{\frac{0.0032}{10}} = 0.0179$.

El coeficiente de determinación es $r^2 = 1 - \frac{0.0032}{0.3019} = 1 - 0.0106 = 0.9894$, esto significa que el modelo de regresión lineal simple explica el 98.94 % de la variabilidad total de las observaciones y_i.

Para construir un intervalo de confianza para la pendiente de nivel $1 - \alpha = 0.95$ buscamos $t_{0.025,10} = 2.2281$ y el intervalo de confianza queda:

$$IC_{0.95}(\beta_1) = 0.0011 \pm 2.2281 \times 0.0179/\sqrt{228800} = (0.00102, 0.00118)$$

■

Si se desea plantear algún test de hipótesis sobre la pendiente β_1, donde la hipótesis nula es que β_1 es un valor fijo b, el estadístico de prueba será:

$$\frac{\widehat{\beta}_1 - b}{s_r/\sqrt{S_{xx}}} \sim T_{(n-2)} \quad \text{si } H_0 \text{ es verdadera}$$

Los diferentes tests que se pueden hacer se resumen como:

RESUMEN 8.1

Bajo el modelo de regresión lineal $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, con $\epsilon_i \sim N(0, \sigma^2)$ e independientes entre sí.			
Hipótesis nula	$H_0 : \beta_1 = b$		
Valor del estadístico de prueba	$t = \frac{\widehat{\beta}_1 - b}{s_r/\sqrt{S_{xx}}}$		
Hipótesis alternativa	$H_A : \beta_1 < b$	$H_A : \beta_1 > b$	$H_A : \beta_1 \neq b$
Región de rechazo con nivel α	$t < -t_{\alpha,n-2}$	$t > t_{\alpha,n-2}$	$\lvert t \rvert > t_{\alpha/2,n-2}$

Ejemplo 8.4

Para los datos del Ejemplo 8.3, si queremos verificar, con nivel de significación $\alpha = 0.05$, que la pendiente $\beta_1 < 0.00115$, debemos plantear:

$$H_0 : \beta_1 = 0.00115 \qquad \text{vs.} \qquad H_A : \beta_1 < 0.00115$$

El estadísitico de prueba es:

$$T = \frac{\widehat{\beta}_1 - 0.00115}{s_r/\sqrt{S_{xx}}} \sim T_{(10)} \text{ si } H_0 \text{ es verdadera}$$

la región de rechazo es:

$$t < -t_{0.05,10} = -1.8125$$

Resolviendo $t = (0.0011 - 0.00115) \times \sqrt{228800}/0.0179 = -4.0085$. Como $-4.0085 < -1.812$ podemos rechazar H_0 con nivel $\alpha = 0.05$, es decir, podemos afirmar que la pendiente es menor que 0.00115 con ese nivel.

■

Intervalo de confianza y test de hipótesis para la ordenada al origen

Si deseamos hacer inferencias sobre β_0 veremos primero cuál es la distribución de su estimador. Para eso veremos que $\widehat{\beta}_0$ se puede escribir como:

$$\begin{aligned}\widehat{\beta}_0 &= \overline{Y} - \overline{x}\widehat{\beta}_1 = \frac{\sum_{i=1}^n Y_i}{n} - \frac{\sum_{i=1}^n (x_i - \overline{x}) Y_i \overline{x}}{S_{xx}} \\ &= \sum_{i=1}^n \left(\frac{1}{n} - \frac{(x_i - \overline{x})\,\overline{x}}{S_{xx}} \right) Y_i\end{aligned}$$

Por lo tanto, $\widehat{\beta}_0$ es una combinación lineal de las Y_i que son independientes y, si tienen distribución Normal, también $\widehat{\beta}_0$ tiene distribución Normal con la media y varianza que ya vimos, es decir:

$$\widehat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)\right) \quad \text{entonces} \quad \frac{\widehat{\beta}_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}} \sim N(0,1)$$

Luego si queremos construir un intervalo de confianza para β_0, la función pivote que debemos usar es:

$$\frac{\widehat{\beta}_0 - \beta_0}{s_r\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}} \sim T_{(n-2)}$$

Siguiendo el procedimiento usual, llegamos al intervalo:

$$IC_{1-\alpha}(\beta_0) = \left(\widehat{\beta}_0 - t_{\alpha/2,n-2} \times s_r \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}, \widehat{\beta}_0 + t_{\alpha/2,n-2} \times s_r \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}} \right)$$



Observación:

La longitud del intervalo para $\widehat{\beta}_0$ es $2t_{\alpha/2,n-2} \times s_r \sqrt{1/n + \overline{x}^2/S_{xx}}$ de modo que, si $\overline{x}$ es relativamente grande, la estimación de β_0 será poco precisa. Generalmente, la estimación de β_0 no es tan importante como la de β_1.

Ejemplo 8.5

Para los datos del Ejemplo 8.3, si elegimos $1 - \alpha = 0.95$, $t_{0.025,10} = 2.2281$ y el intervalo para β_0 es:

$$IC_{0.95}(\beta_0) = 0.0008 \pm 2.2281 \times 0.0179 \times \sqrt{\frac{1}{12} + \frac{300^2}{228800}} = (-0.0267, 0.0283)$$

■

De la misma forma se pueden plantear tests para β_0, pero en general, es de menor interés.

EJERCICIO 8.3

Para calibrar un método analítico, se realizaron determinaciones de magnesio en 6 muestras con concentraciones conocidas. Los valores obtenidos fueron:

Concentración (ppm)	0	2	4	6	8	10
Respuesta	114	870	2087	3353	3970	4950

Hallar un intervalo de confianza para la pendiente de nivel 0.95.

Inferencias en relación a la media y pronóstico de la variable respuesta (Y) correspondiente a un valor x_0 de la variable explicativa

Consideremos x_0 un valor de la variable explicativa y sea Y_0 la respuesta correspondiente. Si se cumple el modelo y x_0 está dentro del rango de las observaciones con las que se estimó la recta, la respuesta media correspondiente a x_0 es:

$$E(Y_0) = \beta_0 + \beta_1 x_0$$

El "valor ajustado":

$$\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$$

se puede considerar un estimador de $E(Y_0)$.

Si deseamos construir un intervalo de confianza para $E(Y_0)$, deberemos encontrar la función pivote adecuada. Es fácil ver que:

$$E(\widehat{y}_0) = E(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) = \beta_0 + \beta_1 x_0$$

y que:

$$V(\widehat{y}_0) = V(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right)$$

También se puede probar que, cuando las Y_i tienen distribución Normal, la función:

$$\frac{\widehat{y}_0 - (\beta_0 + \beta_1 x_0)}{s_r \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}} \sim T_{(n-2)}$$

Entonces, esa es la función pivote que usamos para construir un intervalo de confianza para $E(Y_0)$. Siguiendo el mismo procedimiento de siempre, obtenemos el siguiente intervalo de confianza de nivel $1-\alpha$ para la media de Y, para un valor dado x_0:

$$IC_{1-\alpha}\big(E(Y_0)\big) = \left(\widehat{y}_0 - t_{\alpha/2,n-2} s_r \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}, \widehat{y}_0 + t_{\alpha/2,n-2} s_r \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} \right) \quad (8.5)$$

Ejemplo 8.6

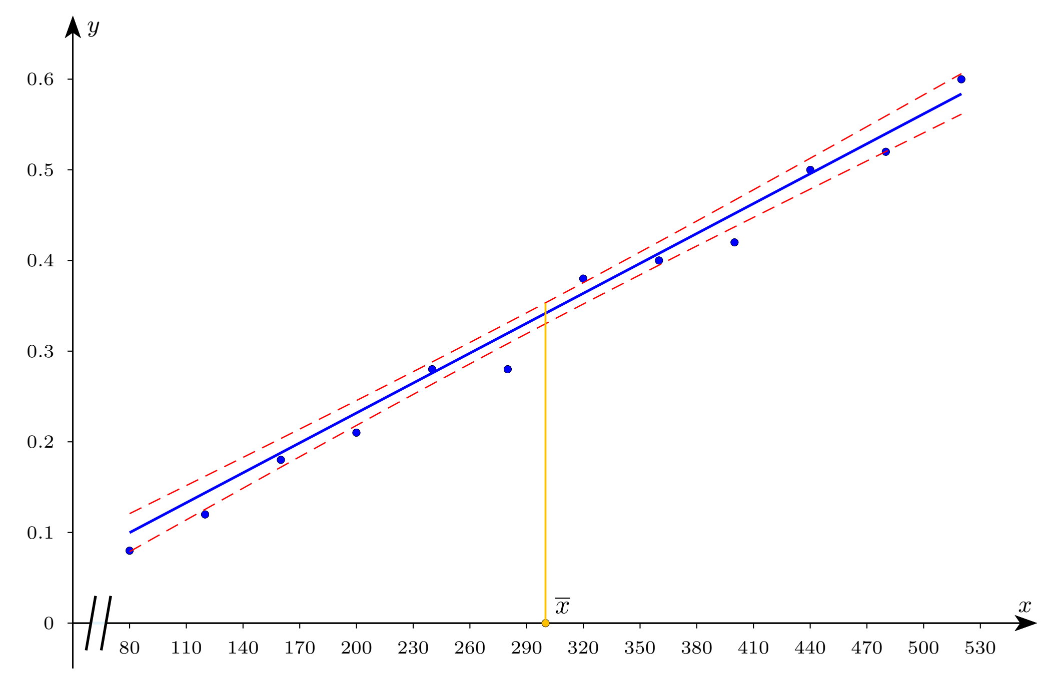
Para los datos del Ejemplo 8.3, si queremos estimar la densidad óptica media que corresponde a una concentración $x_0 = 350$, primero calculamos $\widehat{y}_0 = 0.0008 + 0.0011 \times 350 = 0.3858$ y el $\widehat{dt}(\widehat{y}_0) = \sqrt{\widehat{V}(\widehat{y}_0)} = s_r\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}} = 0.0179 \times \sqrt{\frac{1}{12} + \frac{(350-300)^2}{228800}} = 0.0055$, con el valor crítico $t_{0.025,10} = 2.2281$ el intervalo de confianza es:

$$IC_{0.95}\big(E(Y_0)\big) = 0.3858 \pm 2.2281 \times 0.0055 = (0.3735, 0.3981)$$

A partir del intervalo (8.5), vemos que la longitud es:

$$L = 2t_{\alpha/2,n-2}s_r\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}} \tag{8.6}$$

Esta longitud es mínima cuando x_0 es igual a $\overline{x}$ y aumenta cuando x_0 se aleja de $\overline{x}$. En la siguiente figura, se grafica la recta de regresión estimada (con los datos del Ejemplo 8.6) y dos lineas curvas que representan los límites de los intervalos de confianza para la media de Y, dados los posibles valores de x.



En el gráfico, se puede ver cómo varía la longitud de los intervalos de confianza.



Observación:

Generalmente, el modelo es una aproximación válida, en el mejor de los casos, dentro del rango de las "x" usadas en el experimento. No tenemos información para hacer ninguna inferencia fuera de ese rango de valores, por lo que no es nada confiable "extrapolar", o sea, aplicar este procedimiento para x_0 fuera del rango de las "x" con las que se estimó la recta de regresión.

Si queremos predecir el valor que puede tomar la respuesta, cuando la variable explicativa es x_0, sabemos que $Y_0 = \beta_0 + \beta_1 x_0 + \epsilon$, parece lógico predecir ese valor con el valor sobre la recta estimada, o "valor ajustado":

$$\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$$

éste es el mismo valor que usamos para estimar la $E(Y_0)$. Pero, si pretendemos construir un intervalo de predicción, las cosas cambian un poco. El error de predicción es la diferencia entre el valor que puede tomar una variable aleatoria Y_0 y el valor ajustado $\widehat{y}_0$, podemos ver que el valor esperado del error de predicción es:

$$E(Y_0 - \widehat{y}_0) = 0$$

y la varianza del error de predicción es:

$$V(Y_0 - \widehat{y}_0) = V(Y_0) + V(\widehat{y}_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)$$



Observación:

El valor futuro de Y es independiente de las Y_i observadas anteriormente.

Así, para construir un intervalo de predicción para Y_0 usaremos la función:

$$\frac{Y_0 - \widehat{y}_0}{s_r\sqrt{1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}}} \sim T_{(n-2)}$$

y el intervalo de predicción para Y_0 es:

$$IP_{1-\alpha}(Y_0) = \widehat{y}_0 \pm t_{\alpha/2,n-2}\ s_r\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} \tag{8.7}$$

Ejemplo 8.7

Con los datos del Ejemplo 8.3, si queremos hacer un intervalo de predicción para un posible valor de la densidad óptica correspondiente a la misma concentración $x_0 = 350$, el valor $\widehat{y}_0$ es el mismo, pero $dt(\widehat{Y_0 - \widehat{y}_0}) = \sqrt{\widehat{V(Y_0 - \widehat{y}_0)}} = s_r\sqrt{1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}} = 0.0179 \times \sqrt{1 + \frac{1}{12} + \frac{(350-300)^2}{228800}} = 0.0187$. Luego, el intervalo de predicción será:

$$IP_{0.95}(Y_0) = 0.3858 \pm 2.2281 \times 0.0187 = (0.3441, 0.4275)$$

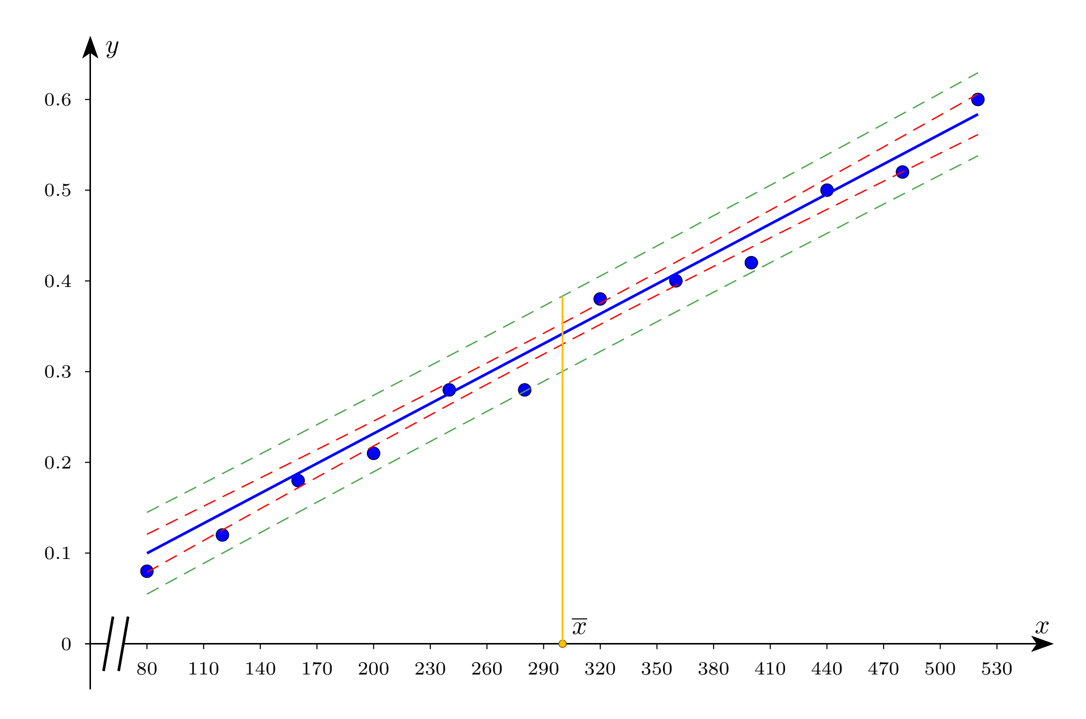
esto significa que tenemos un 95 % de confianza de que ese intervalo contenga a la posible respuesta Y_0 correspondiente a una concentración $x_0 = 350$.

■

La longitud del intervalo de predicción para Y_0 (8.7) es

$$L = 2t_{\alpha/2,n-2}\ s_r\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$$

Vale lo mismo que dijimos para los intervalos de confianza, la longitud es mínima cuando x_0 es igual a $\overline{x}$. Además, la longitud de este intervalo es mayor que la del intervalo de confianza para $E(Y_0)$ (8.6). Ésto es lógico porque, para predecir el valor que tome la variable aleatoria tengo más incertidumbre que para estimar su media. Podemos visualizar ésto con los datos del ejemplo anterior. En la siguiente figura se observan los puntos (x_i, y_i), la recta de regresión estimada (línea sólida de color azul), dos pares de curvas que representan los límites de los intervalos de confianza (líneas de rayas de color rojo) y los intervalos de predicción (líneas de rayas de color verde).



EJERCICIO 8.4

Para los datos del Ejercicio 8.3 construir, si es posible, un intervalo para la respuesta media y de predicción de nivel 0.95 para $x = 3, 7$ y 12.

Curva de calibración

Un ejemplo particular del uso de un modelo de regresión lineal es construir una curva de calibración. Una vez establecida la curva de calibrado, ésta puede usarse para obtener la concentración del analito, correspondiente a cualquier material de ensayo. Por este motivo, es importante que los patrones que se usan para construir la curva cubran el intervalo completo de concentraciones requerido en subsiguientes análisis.

El procedimiento habitual es el siguiente: el analista toma una serie de materiales con concentraciones del analito conocidas $x_1, x_2, \ldots, x_n$. Estos patrones de calibración se miden en el instrumento analítico bajo las mismas condiciones que las utilizadas posteriormente

para los materiales de ensayo (es decir, los "desconocidos"), obteniéndose los valores $y_1, y_2, \ldots, y_n$. Luego, se grafican los n pares $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ y se observa si están próximos a una curva.

La primera cuestión que se plantea es si la curva de calibración es lineal. En este curso solamente analizaremos el caso de una recta, pero hay procedimientos similares para ajustar otro tipo de curvas.

En el caso de observar linealidad, es válido asumir el modelo de regresión lineal:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

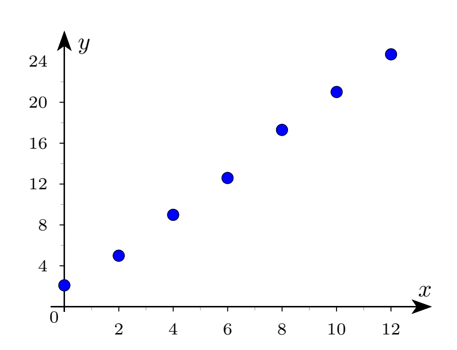
donde β_0 y β_1 son parámetros fijos y los ϵ_i son errores de medición aleatorios independientes entre sí, y en consecuencia, podemos suponer que tienen distribución normal, $\epsilon_i \sim N(0, \sigma^2)$.

Ejemplo 8.8

Para construir una curva de calibración, se han examinado una serie de soluciones patrones de fluorescencia en un espectrómetro de fluorescencia y han conducido a las siguientes intensidades de fluorescencia (en unidades arbitrarias):

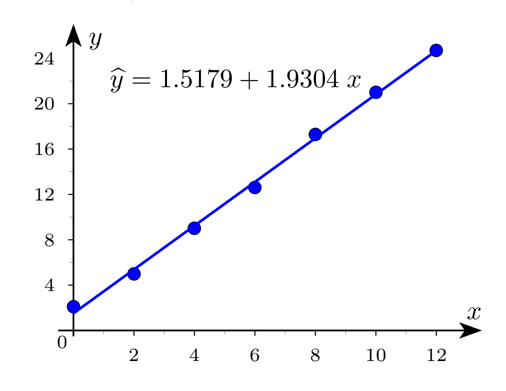
intensidad de fluorescencia	2.1	5.0	9.0	12.6	17.3	21.0	24.7
concentración *pg ml*$^{-1}$	0	2	4	6	8	10	12

Si graficamos estos valores:



vemos que los puntos parecen estar bastante próximos a una recta y podemos aceptar que la relación entre las variables es "aproximadamente lineal". Podemos estimar los parámetros desconocidos β_0, β_1 y σ^2.

Tenemos que: $S_{xx} = 112$, $S_{xy} = 216.2$, $S_{yy} = 418.28$, $S_{rr} = 0.9368$, $\overline{x} = 6$ y $\overline{y} = 13.1$. Luego, obtenemos $\widehat{\beta}_0 = 1.5179$ y $\widehat{\beta}_1 = 1.9304$, así la recta estimada es:



Si queremos construir un intervalo de confianza para la pendiente, debemos calcular $\widehat{dt(\widehat{\beta}_1)} = s_r/\sqrt{S_{xx}} = 0.4329/\sqrt{112} = 0.0409$ eligiendo $1-\alpha = 0.95$ tenemos $t_{0.025,5} = 2.5706$ y el intervalo para β_1 es:

$$IC_{0.95}(\beta_1) = 1.9304 \pm 2.5706 \times 0.0409 = (1.8253, 2.0355)$$

Ahora, si elegimos $x_0 = 5$, se tiene que $\widehat{y}_0 = 1.5179 + 1.9304 \times 5 = 11.1699$, $\widehat{dt(\widehat{y}_0)} = 0.4329 \times \sqrt{\frac{1}{7} + \frac{(6-5)^2}{112}} = 0.1687$ y el intervalo de 95 % de confianza para $E(Y_0)$ resulta:

$$IC_{0.95}\big(E(Y_0)\big) = 11.1699 \pm 2.5706 \times 0.1687 = (10.7362, 11.6036),$$

esto significa que tenemos un 95 % de confianza de que este intervalo contenga el valor verdadero (desconocido) de $E(Y_0)$, que es el valor medio de las respuestas correspondientes a la concentración $x_0 = 5$.

También se puede construir un intervalo de predicción, por ejemplo, para $x = 9$ obtendremos:

$$IP_{0.95}(Y_0) = 18.8915 \pm 2.5706 \times 0.2045 = (18.3658, 19.4172)$$

■

Estimación de un valor de x a partir de un valor de y

La curva de calibración, luego de analizar si se ajusta al modelo lineal, se usa para conocer la verdadera concentración (x_0) correspondiente a una señal medida en el instrumento (y_0), ahora bien, si se tratara de un modelo determinístico, se podría despejar x_0 de la ecuación:

$$y_0 = \beta_0 + \beta_1 x_0$$

resultando:

$$x_0 = \frac{y_0 - \beta_0}{\beta_1}$$

Pero estamos trabajando con una recta estimada:

$$\widehat{y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$$

donde $\widehat{y}_0$, $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son variables aleatorias. Entonces, estimamos el valor x_0 como:

$$\begin{aligned}\widehat{x}_0 &= \frac{\widehat{y}_0 - \widehat{\beta}_0}{\widehat{\beta}_1} = \frac{\widehat{y}_0 - \left(\overline{Y} - \overline{x}\widehat{\beta}_1\right)}{\widehat{\beta}_1} = \frac{\widehat{y}_0 - \overline{Y} + \overline{x}\widehat{\beta}_1}{\widehat{\beta}_1} \\ &= \frac{\widehat{y}_0 - \overline{Y}}{\widehat{\beta}_1} + \overline{x}\end{aligned}$$

Luego $\widehat{x}_0$ es una variable aleatoria, ya que es función de $\widehat{y}_0$, $\overline{Y}$ y $\widehat{\beta}_1$, que son variables aleatorias.

Si queremos tener una idea de la precisión de esta estimación, recordando el tema de propagación de errores (visto en el Capítulo 4), podemos aproximar la desviación estándar de $\widehat{x}_0$:

$$dt\left(\widehat{x}_0\right) \cong \frac{\sigma}{\beta_1}\sqrt{1 + \frac{1}{n} + \frac{(y_0 - E(\overline{Y}))^2}{\beta_1^2 S_{xx}}}$$

y podemos estimarla con:

$$\widehat{dt\left(\widehat{x}_0\right)} \cong \frac{s_r}{\widehat{\beta}_1}\sqrt{1+\frac{1}{n}+\frac{(y_0-\overline{y})^2}{\widehat{\beta}_1^2 S_{xx}}}$$



Observación:

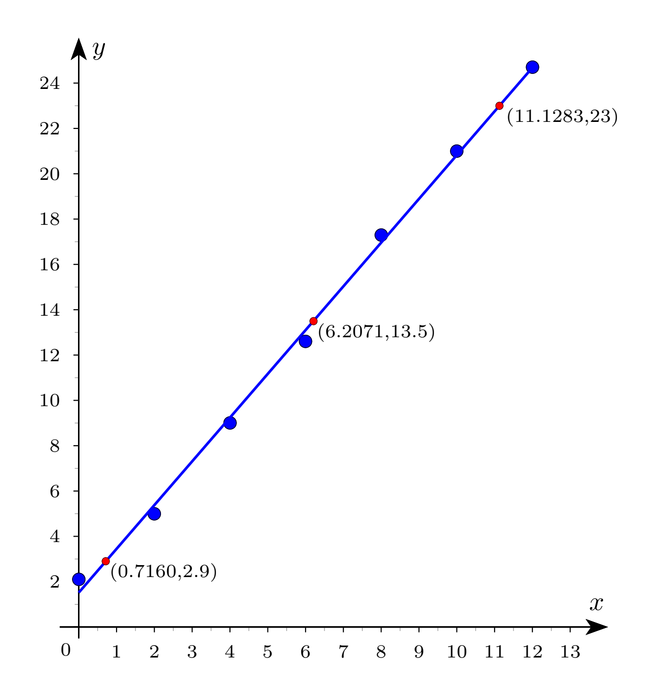
A partir de la fórmula, se puede observar que la desviación estándar es menor cuanto más cerca está el valor y_0 del valor $\overline{y}$ calculado con los datos usados para construir la curva de calibración.

Ejemplo 8.9

Volvamos con los datos del Ejemplo 8.8. Determinaremos para los siguientes tres valores de fluorescencia: 2.9, 13.5 y 23, su concentración estimada y su desviación estándar aproximada, que se muestran a continuación:

Valor de fluorescencia	concentración estimada	desviación típica aproximada
$y_0 = 2.9$	$\widehat{x}_0 = \frac{2.9-1.5129}{1.9304} = 0.7160$	$\frac{0.4329}{1.9304}\sqrt{1+\frac{1}{7}+\frac{(2.9-13.1)^2}{1.9304^2\times 112}} = 0.2646$
$y_0 = 13.5$	$\widehat{x}_0 = \frac{13.5-1.5129}{1.9304} = 6.2071$	$\frac{0.4329}{1.9304}\sqrt{1+\frac{1}{7}+\frac{(13.5-13.1)^2}{1.9304^2\times 112}} = 0.2398$
$y_0 = 23$	$\widehat{x}_0 = \frac{23.0-1.5129}{1.9304} = 11.1283$	$\frac{0.4329}{1.9304}\sqrt{1+\frac{1}{7}+\frac{(23.0-13.1)^2}{1.9304^2\times 112}} = 0.2632$

Gráficamente



EJERCICIO 8.5

La respuesta de un ensayo colorimétrico para glucosa se controló con la ayuda de soluciones patrón de glucosa. A partir de la siguiente muestra:

Glucosa (nM)	0	2	4	6	8	10
Absorbancia	0.002	0.150	0.294	0.434	0.570	0.704

Hallar, si es posible, para los siguientes valores de absorbancia: 0.3, 0.5 y 0.9, los valores estimados de glucosa y sus respectivos errores estándar.

Referencias

Agresti, A. & Franklin, C. A. (2009). *Statistics: The Art and Science of learning from Data.* Pearson New International edition.

Altman, D. G. (1990). *Practical Statistics for Medical Research.* Published by Chapman & Hall.

Daniel, W. (2002). *Bioestadística: Base para el análisis de las ciencias de la salud.* Ed. Limusa Wiley.

Devore Jay, L. (2001). *Probabilidad y Estadística para Ingeniería y Ciencias.* Ed. Books/Cole Publishing Company.

Dixon, W. & Massey, F. (1970). *Introducción al Análisis Estadístico.* México. Libros Mc Graw-Hill.

Maronna, R. (1995). *Probabilidad y Estadística Elementales para Estudiantes de Ciencias.* Buenos Aires.Ed. Exactas.

Mendenhall, W., Beaver, R. J. & Beaver, B. M. (2006). *Introducción a la Probabilidad y Estadística.* México. Cengage Learning Editores.

Ross, S. M. (1987). *Introduction to Probability and Statistics for Enginees and Scientists.* Published by John Wiley & Sons.

Wackerly, D. D., Mendenhall, W. & Scheaffer, R. L. (2010). *Estadística Matemática con aplicaciones.* México. Cengage Learning Editores. Walpole, R. E. & Myers, R. H. (2007). *Probabilidad y Estadística para Ingeniería y Ciencias.* México. Ediciones McGraw-Hill.

APÉNDICE A

Teoría de Conjuntos

Aquí no se darán las explicaciones teóricas de la Teoría de Conjuntos, pues se pueden encontrar en cualquier texto básico de Álgebra, sino que se hará un repaso de los conceptos básicos que se requieren, algo así como un "ayuda memoria".

A continuación se repasarán las definiciones y propiedades básicas:

Definición:

Dado un conjunto $A = \{a, b, c, d\}$, la **pertenencia** del elemento a al conjunto A se representa por $a \in A$.

Definición:

Se llama **cardinal** del conjunto al número de elementos que contiene, notamos $\# A$.

Definición:

Se llama **conjunto vacío**, y se representa por $\emptyset$, al conjunto que no contienen ningún elemento.

Definición:

Se llama **universo** o conjunto universal, y se suele representar por Ω, al conjunto formado por todos los elementos que se están considerando.

Definición:

Dado un conjunto A, se llama **complemento** del mismo, y se representa por A^c o A', al conjunto formado por los elemntos del universo que no están en A.

Aclaración

Dos conjuntos son iguales si están formados por los mismos elementos.

Definición:

Se dice que B es **subconjunto** de A, y se representa $B \subseteq A$, si todos los elementos de B pertenecen a A. Se dice también que B está incluído en A.

Definición:

Dados dos conjuntos A y B, se llama **unión** de ambos, y se representa $A \cup B$, al conjunto formado por los elmentos que pertenecen a A o a B.

Definición:

Dados dos conjuntos A y B, se llama **intersección** de ambos y se representa $A \cap B$, al conjunto formado por los elementos que pertenecen a A y a B.

Definición:

Si dos conjuntos no tienen elementos comunes, se llaman **disjuntos**, **incompatibles** o **mutuamente excluyentes** y su intercción es el conjunto vacío.

Definición:

Dados dos conjuntos, A y B, la **diferencia** entre A y B es el conjunto de los elementos que están en A y no están en B y se representa por $A - B$.

A continuación repasaremos algunas propiedades básicas:

PROPIEDADES DE LA INCLUSIÓN:

- $\emptyset \subseteq A \subseteq \Omega$
- *Reflexiva:* $A \subseteq A$
- *Antisimétrica:* $A \subseteq B$ y $B \subseteq A$ implica $A = B$
- *Transitiva:* $A \subseteq B$ y $B \subseteq D$ implica $A \subseteq D$
- $A \cap B \subseteq A$ y $A \cap B \subseteq B$
- $A \cap B = A$ entonces $A \subseteq B$
- $A \cap B \subseteq A \cup B$

PROPIEDADES DE LA UNIÓN E INTERSECCIÓN:

- *Identidad:* $A \cup \emptyset = A$, $A \cup \Omega = \Omega$, $A \cap \Omega = A$ y $A \cap \emptyset = \emptyset$
- *Idempotencia:* $A \cup A = A$ y $A \cap A = A$
- *Conmutatividad:* $A \cup B = B \cup A$ y $A \cap B = B \cap A$
- *Asociatividad:* $(A \cup B) \cup D = A \cup (B \cup D)$ y $(A \cap B) \cap D = A \cap (B \cap D)$
- *Distributividad:* $(A \cup B) \cap D = (A \cap D) \cup (B \cap D)$ y $(A \cap B) \cup D = (A \cup D) \cap (B \cup D)$
- *Absorción:* $A \cup (A \cap B) = A$ y $A \cap (A \cup B) = A$
- *Complementaridad:* $A \cup A^c = \Omega$ y $A \cap A^c = \emptyset$

PROPIEDADES DEL COMPLEMENTO:

- $(A^c)^c = A$, $\emptyset^c = \Omega$ y $\Omega^c = \emptyset$
- $A \subseteq B$ entonces $B^c \subseteq A^c$
- Leyes de De Morgan: $(A \cup B)^c = A^c \cap B^c$ y $(A \cap B)^c = A^c \cup B^c$
- $A - B = A \cap B^c = A - (A \cap B)$

PROPIEDADES DE LA DIFERENCIA:

- $A - A = \emptyset$, $A - \emptyset = A$, $A - \Omega = \emptyset$, $\emptyset - A = \emptyset$ y $\Omega - A = A^c$
- $A - B \subseteq A$
- $A - B = A$ entonces $A \cap B = \emptyset$
- $(A - B) - D = (A - D) - B$
- $A \cap (B - D) = (A \cap B) - (A \cap D)$

APÉNDICE B

Tablas

Aquí se encuentran las Tablas de algunas distribuciones utilizadas en este libro. Los resultados fueron redondeados a 4 decimales. A continuación se dará una breve descripción del uso de las mismas:

- **Probabilidades acumulativas Binomial**: se tabulan los valores de la fda de la v.a. $X \sim B(n,p)$. Hay una tabla para cada uno de los siguientes valores de $n = 5$, 10, 15, 20 y 25, y en cada una, los valores de p son: 0.01, 0.05, 0.1, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.9, 0.95 y 0.99, que se encuentran en la primera fila. En la primera columna, titulada k, se encuentran los posibles valores de la v.a. en este caso k es un valor natural entre 0 y $n-1$ (es claro que $F(n) = 1$, por eso este valor se omite). En las restantes columnas se tabulan los valores de $F(k)$ correspondientes a los distintos valores de p. Por ejemplo, si $X \sim B(5, 0.75)$ obtenemos que $F(3) = 0.3672$ como se muestra la siguiente gráfica que corresponde a un fragmento de la original.

Binomial con $n = 5$

k	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
0	0.9510	0.7738	0.5905	0.3277	0.2373	0.1681	0.0778	0.0313	0.0102	0.0024	0.0010	0.0003	0.0000	0.0000	0.0000
1	0.9990	0.9774	0.9185	0.7373	0.6328	0.5282	0.3370	0.1875	0.0870	0.0308	0.0156	0.0067	0.0005	0.0000	0.0000
2	1.0000	0.9988	0.9914	0.9421	0.8965	0.8369	0.6826	0.5000	0.3174	0.1631	0.1035	0.0579	0.0086	0.0012	0.0000
3	1.0000	1.0000	0.9995	0.9933	0.9844	0.9692	0.9130	0.8125	0.6630	0.4718	**0.3672**	0.2627	0.0815	0.0226	0.0010
4	1.0000	1.0000	1.0000	0.9997	0.9990	0.9976	0.9898	0.9688	0.9222	0.8319	0.7627	0.6723	0.4095	0.2262	0.0490

- **Probabilidades acumulativas de Poisson**: se tabulan los valores de la fda de la v.a. $X \sim P(\lambda)$. En la primera fila se encuentran los siguientes valores de λ: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 17, 20 y 25. En la primera columna se encuentran los posibles valores de la v.a., en este caso $k = 0, 1, 2, \ldots$. En las restantes columnas se tabulan los valores de $F(k)$ correspondientes a los distintos valores de λ. Por ejemplo, si $X \sim P(0.9)$, tenemos que que $F(5) = 0.9997$ (como se muestra la siguiente gráfica

que corresponde a un fragmento de la original).

Poisson													
k	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**	**2**	**3**	**4**
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679	0.1353	0.0498	0.0183
1	0.9953	0.9825	0.9631	0.9384	0.9098	0.8781	0.8442	0.8088	0.7725	0.7358	0.4060	0.1991	0.0916
2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9769	0.9659	0.9526	0.9371	0.9197	0.6767	0.4232	0.2381
3	1.0000	0.9999	0.9997	0.9992	0.9982	0.9966	0.9942	0.9909	0.9865	0.9810	0.8571	0.6472	0.4335
4		1.0000	1.0000	0.9999	0.9998	0.9996	0.9992	0.9986	0.9977	0.9963	0.9473	0.8153	0.6288
5				1.0000	1.0000	1.0000	0.9999	0.9998	**0.9997**	0.9994	0.9834	0.9161	0.7851
6							1.0000	1.0000	1.0000	0.9999	0.9955	0.9665	0.8893

- **Función de Distribución Normal Típica**: se tabulan los valores de $\Phi(z) = P(Z \leq z)$ para la v.a. $Z \sim N(0,1)$ y los valores disponibles de z son los números reales entre -3.59 y 3.59 expresados con dos decimales. En la primer columna se indica la parte entera y el primer decimal y en la primer fila el segundo decimal del número z. Por ejemplo, si necesitamos el valor de $\Phi(-2.92) = 0.0018$ (como se muestra en la siguiente gráfica que es un fragmento de la original).

Normal estándar										
z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	**0.0018**	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019

- **Valores críticos para la distribución Student**: se proporcionan los valores críticos $t_{\alpha,n}$ donde n son los grados de libertad (g.l.). Los g.l. disponibles son: del 1 al 35, 40, 50, 60, 100, 120 y muy grande (∞), indicados en la primera columna, y los valores de α son: 0.1, 0.08, 0.05, 0.025, 0.01, 0.005, 0.0025, 0.001, 0.0008 y 0.0005, indicados en la primera fila. Por ejemplo, si necesitamos averiguar el valor crítico de $t_{0.0025,4}$, obtendremos el valor 5.5976 como se muestra a continuación:

Student										
g.l.	**0.1000**	**0.0800**	**0.0500**	**0.0250**	**0.0100**	**0.0050**	**0.0025**	**0.0010**	**0.0008**	**0.0005**
1	3.0777	3.8947	6.3138	12.7062	31.8205	63.6567	127.3213	318.3088	397.8865	636.6192
2	1.8856	2.1894	2.9200	4.3027	6.9646	9.9248	14.0890	22.3271	24.9700	31.5991
3	1.6377	1.8589	2.3534	3.1824	4.5407	5.8409	7.4533	10.2145	11.0207	12.9240
4	1.5332	1.7229	2.1318	2.7764	3.7469	4.6041	**5.5976**	7.1732	7.6104	8.6103
5	1.4759	1.6493	2.0150	2.5706	3.3649	4.0321	4.7733	5.8934	6.1943	6.8688

- **Valores críticos para la distribución Chi-cuadrado**: se proporcionan los valores críticos $\chi^2_{\alpha,n}$. Los g.l. disponibles son desde el 1 hasta el 45, indicados en la primera columna, y los valores de α son: 0.0025, 0.0050, 0.0100, 0.0250, 0.0500, 0.9500, 0.9750, 0.9900, 0.9950 y 0.9975, indicados en la primera fila. Por ejemplo, si necesitamos buscar el valor crítico de $\chi^2_{0.9750,4}$ con g.l.=4, obtendremos el valor 0.4844 (como se muestra en la siguiente gráfica que es un fragmento de la original).

Chi-cuadrado										
g.l.	**0.9975**	**0.9950**	**0.9900**	**0.9750**	**0.9500**	**0.0500**	**0.0250**	**0.0100**	**0.0050**	**0.0025**
1	0.0000	0.0000	0.0002	0.0010	0.0039	3.8415	5.0239	6.6349	7.8794	9.1406
2	0.0050	0.0100	0.0201	0.0506	0.1026	5.9915	7.3778	9.2103	10.5966	11.9829
3	0.0449	0.0717	0.1148	0.2158	0.3518	7.8147	9.3484	11.3449	12.8382	14.3203
4	0.1449	0.2070	0.2971	**0.4844**	0.7107	9.4877	11.1433	13.2767	14.8603	16.4239
5	0.3075	0.4117	0.5543	0.8312	1.1455	11.0705	12.8325	15.0863	16.7496	18.3856
6	0.5266	0.6757	0.8721	1.2373	1.6354	12.5916	14.4494	16.8119	18.5476	20.2494

Binomial con $n = 5$															
k	**0.01**	**0.05**	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.75**	**0.80**	**0.90**	**0.95**	**0.99**
0	0.9510	0.7738	0.5905	0.3277	0.2373	0.1681	0.0778	0.0313	0.0102	0.0024	0.0010	0.0003	0.0000	0.0000	0.0000
1	0.9990	0.9774	0.9185	0.7373	0.6328	0.5282	0.3370	0.1875	0.0870	0.0308	0.0156	0.0067	0.0005	0.0000	0.0000
2	1.0000	0.9988	0.9914	0.9421	0.8965	0.8369	0.6826	0.5000	0.3174	0.1631	0.1035	0.0579	0.0086	0.0012	0.0000
3	1.0000	1.0000	0.9995	0.9933	0.9844	0.9692	0.9130	0.8125	0.6630	0.4718	0.3672	0.2627	0.0815	0.0226	0.0010
4	1.0000	1.0000	1.0000	0.9997	0.9990	0.9976	0.9898	0.9688	0.9222	0.8319	0.7627	0.6723	0.4095	0.2262	0.0490

Binomial con $n = 10$															
k	**0.01**	**0.05**	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.75**	**0.80**	**0.90**	**0.95**	**0.99**
0	0.9044	0.5987	0.3487	0.1074	0.0563	0.0282	0.0060	0.0010	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.9957	0.9139	0.7361	0.3758	0.2440	0.1493	0.0464	0.0107	0.0017	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.9999	0.9885	0.9298	0.6778	0.5256	0.3828	0.1673	0.0547	0.0123	0.0016	0.0004	0.0001	0.0000	0.0000	0.0000
3	1.0000	0.9990	0.9872	0.8791	0.7759	0.6496	0.3823	0.1719	0.0548	0.0106	0.0035	0.0009	0.0000	0.0000	0.0000
4	1.0000	0.9999	0.9984	0.9672	0.9219	0.8497	0.6331	0.3770	0.1662	0.0473	0.0197	0.0064	0.0001	0.0000	0.0000
5	1.0000	1.0000	0.9999	0.9936	0.9803	0.9527	0.8338	0.6230	0.3669	0.1503	0.0781	0.0328	0.0016	0.0001	0.0000
6	1.0000	1.0000	1.0000	0.9991	0.9965	0.9894	0.9452	0.8281	0.6177	0.3504	0.2241	0.1209	0.0128	0.0010	0.0000
7	1.0000	1.0000	1.0000	0.9999	0.9996	0.9984	0.9877	0.9453	0.8327	0.6172	0.4744	0.3222	0.0702	0.0115	0.0001
8	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9983	0.9893	0.9536	0.8507	0.7560	0.6242	0.2639	0.0861	0.0043
9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9990	0.9940	0.9718	0.9437	0.8926	0.6513	0.4013	0.0956

Binomial con $n = 15$															
k	**0.01**	**0.05**	**0.10**	**0.20**	**0.25**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.75**	**0.80**	**0.90**	**0.95**	**0.99**
0	0.8601	0.4633	0.2059	0.0352	0.0134	0.0047	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.9904	0.8290	0.5490	0.1671	0.0802	0.0353	0.0052	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.9996	0.9638	0.8159	0.3980	0.2361	0.1268	0.0271	0.0037	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	1.0000	0.9945	0.9444	0.6482	0.4613	0.2969	0.0905	0.0176	0.0019	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
4	1.0000	0.9994	0.9873	0.8358	0.6865	0.5155	0.2173	0.0592	0.0093	0.0007	0.0001	0.0000	0.0000	0.0000	0.0000
5	1.0000	0.9999	0.9978	0.9389	0.8516	0.7216	0.4032	0.1509	0.0338	0.0037	0.0008	0.0001	0.0000	0.0000	0.0000
6	1.0000	1.0000	0.9997	0.9819	0.9434	0.8689	0.6098	0.3036	0.0950	0.0152	0.0042	0.0008	0.0000	0.0000	0.0000
7	1.0000	1.0000	1.0000	0.9958	0.9827	0.9500	0.7869	0.5000	0.2131	0.0500	0.0173	0.0042	0.0000	0.0000	0.0000
8	1.0000	1.0000	1.0000	0.9992	0.9958	0.9848	0.9050	0.6964	0.3902	0.1311	0.0566	0.0181	0.0003	0.0000	0.0000
9	1.0000	1.0000	1.0000	0.9999	0.9992	0.9963	0.9662	0.8491	0.5968	0.2784	0.1484	0.0611	0.0022	0.0001	0.0000
10	1.0000	1.0000	1.0000	1.0000	0.9999	0.9993	0.9907	0.9408	0.7827	0.4845	0.3135	0.1642	0.0127	0.0006	0.0000
11	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9981	0.9824	0.9095	0.7031	0.5387	0.3518	0.0556	0.0055	0.0000
12	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9963	0.9729	0.8732	0.7639	0.6020	0.1841	0.0362	0.0004
13	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9948	0.9647	0.9198	0.8329	0.4510	0.1710	0.0096
14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9953	0.9866	0.9648	0.7941	0.5367	0.1399

Binomial con $n = 20$

k	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
0	0.8179	0.3585	0.1216	0.0115	0.0032	0.0008	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.9831	0.7358	0.3917	0.0692	0.0243	0.0076	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.9990	0.9245	0.6769	0.2061	0.0913	0.0355	0.0036	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	1.0000	0.9841	0.8670	0.4114	0.2252	0.1071	0.0160	0.0013	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	1.0000	0.9974	0.9568	0.6296	0.4148	0.2375	0.0510	0.0059	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	1.0000	0.9997	0.9887	0.8042	0.6172	0.4164	0.1256	0.0207	0.0016	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	1.0000	1.0000	0.9976	0.9133	0.7858	0.6080	0.2500	0.0577	0.0065	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000
7	1.0000	1.0000	0.9996	0.9679	0.8982	0.7723	0.4159	0.1316	0.0210	0.0013	0.0002	0.0000	0.0000	0.0000	0.0000
8	1.0000	1.0000	0.9999	0.9900	0.9591	0.8867	0.5956	0.2517	0.0565	0.0051	0.0009	0.0001	0.0000	0.0000	0.0000
9	1.0000	1.0000	1.0000	0.9974	0.9861	0.9520	0.7553	0.4119	0.1275	0.0171	0.0039	0.0006	0.0000	0.0000	0.0000
10	1.0000	1.0000	1.0000	0.9994	0.9961	0.9829	0.8725	0.5881	0.2447	0.0480	0.0139	0.0026	0.0000	0.0000	0.0000
11	1.0000	1.0000	1.0000	0.9999	0.9991	0.9949	0.9435	0.7483	0.4044	0.1133	0.0409	0.0100	0.0001	0.0000	0.0000
12	1.0000	1.0000	1.0000	1.0000	0.9998	0.9987	0.9790	0.8684	0.5841	0.2277	0.1018	0.0321	0.0004	0.0000	0.0000
13	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9935	0.9423	0.7500	0.3920	0.2142	0.0867	0.0024	0.0000	0.0000
14	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9984	0.9793	0.8744	0.5836	0.3828	0.1958	0.0113	0.0003	0.0000
15	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9941	0.9490	0.7625	0.5852	0.3704	0.0432	0.0026	0.0000
16	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9987	0.9840	0.8929	0.7748	0.5886	0.1330	0.0159	0.0000
17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9964	0.9645	0.9087	0.7939	0.3231	0.0755	0.0010
18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9924	0.9757	0.9308	0.6083	0.2642	0.0169
19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9992	0.9968	0.9885	0.8784	0.6415	0.1821

Binomial con $n = 25$

k	0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
0	0.7778	0.2774	0.0718	0.0038	0.0008	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.9742	0.6424	0.2712	0.0274	0.0070	0.0016	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.9980	0.8729	0.5371	0.0982	0.0321	0.0090	0.0004	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.9999	0.9659	0.7636	0.2340	0.0962	0.0332	0.0024	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	1.0000	0.9928	0.9020	0.4207	0.2137	0.0905	0.0095	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	1.0000	0.9988	0.9666	0.6167	0.3783	0.1935	0.0294	0.0020	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
6	1.0000	0.9998	0.9905	0.7800	0.5611	0.3407	0.0736	0.0073	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
7	1.0000	1.0000	0.9977	0.8909	0.7265	0.5118	0.1536	0.0216	0.0012	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
8	1.0000	1.0000	0.9995	0.9532	0.8506	0.6769	0.2735	0.0539	0.0043	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
9	1.0000	1.0000	0.9999	0.9827	0.9287	0.8106	0.4246	0.1148	0.0132	0.0005	0.0000	0.0000	0.0000	0.0000	0.0000
10	1.0000	1.0000	1.0000	0.9944	0.9703	0.9022	0.5858	0.2122	0.0344	0.0018	0.0002	0.0000	0.0000	0.0000	0.0000
11	1.0000	1.0000	1.0000	0.9985	0.9893	0.9558	0.7323	0.3450	0.0778	0.0060	0.0009	0.0001	0.0000	0.0000	0.0000
12	1.0000	1.0000	1.0000	0.9996	0.9966	0.9825	0.8462	0.5000	0.1538	0.0175	0.0034	0.0004	0.0000	0.0000	0.0000
13	1.0000	1.0000	1.0000	0.9999	0.9991	0.9940	0.9222	0.6550	0.2677	0.0442	0.0107	0.0015	0.0000	0.0000	0.0000
14	1.0000	1.0000	1.0000	1.0000	0.9998	0.9982	0.9656	0.7878	0.4142	0.0978	0.0297	0.0056	0.0000	0.0000	0.0000
15	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9868	0.8852	0.5754	0.1894	0.0713	0.0173	0.0001	0.0000	0.0000
16	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9957	0.9461	0.7265	0.3231	0.1494	0.0468	0.0005	0.0000	0.0000
17	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9988	0.9784	0.8464	0.4882	0.2735	0.1091	0.0023	0.0000	0.0000
18	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9927	0.9264	0.6593	0.4389	0.2200	0.0095	0.0002	0.0000
19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9980	0.9706	0.8065	0.6217	0.3833	0.0334	0.0012	0.0000
20	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9995	0.9905	0.9095	0.7863	0.5793	0.0980	0.0072	0.0000
21	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9976	0.9668	0.9038	0.7660	0.2364	0.0341	0.0001
22	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9996	0.9910	0.9679	0.9018	0.4629	0.1271	0.0020
23	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9984	0.9930	0.9726	0.7288	0.3576	0.0258
24	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9992	0.9962	0.9282	0.7226	0.2222

Poisson

k	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	2	3	4	5	6	7	8	9	10	12	15	17	20	25
0	0.9048	0.8187	0.7408	0.6703	0.6065	0.5488	0.4966	0.4493	0.4066	0.3679	0.1353	0.0498	0.0183	0.0067	0.0025	0.0009	0.0003	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.9953	0.9825	0.9631	0.9384	0.9098	0.8781	0.8442	0.8088	0.7725	0.7358	0.4060	0.1991	0.0916	0.0404	0.0174	0.0073	0.0030	0.0012	0.0005	0.0001	0.0000	0.0000	0.0000	0.0000
2	0.9998	0.9989	0.9964	0.9921	0.9856	0.9769	0.9659	0.9526	0.9371	0.9197	0.6767	0.4232	0.2381	0.1247	0.0620	0.0296	0.0138	0.0062	0.0028	0.0005	0.0000	0.0000	0.0000	0.0000
3	1.0000	0.9999	0.9997	0.9992	0.9982	0.9966	0.9942	0.9909	0.9865	0.9810	0.8571	0.6472	0.4335	0.2650	0.1512	0.0818	0.0424	0.0212	0.0103	0.0023	0.0002	0.0000	0.0000	0.0000
4		1.0000	1.0000	0.9999	0.9998	0.9996	0.9992	0.9986	0.9977	0.9963	0.9473	0.8153	0.6288	0.4405	0.2851	0.1730	0.0996	0.0550	0.0293	0.0076	0.0009	0.0002	0.0000	0.0000
5				1.0000	1.0000	1.0000	0.9999	0.9998	0.9997	0.9994	0.9834	0.9161	0.7851	0.6160	0.4457	0.3007	0.1912	0.1157	0.0671	0.0203	0.0028	0.0007	0.0001	0.0000
6							1.0000	1.0000	1.0000	0.9999	0.9955	0.9665	0.8893	0.7622	0.6063	0.4497	0.3134	0.2068	0.1301	0.0458	0.0076	0.0021	0.0003	0.0000
7										1.0000	0.9989	0.9881	0.9489	0.8666	0.7440	0.5987	0.4530	0.3239	0.2202	0.0895	0.0180	0.0054	0.0008	0.0000
8											0.9998	0.9962	0.9786	0.9319	0.8472	0.7291	0.5925	0.4557	0.3328	0.1550	0.0374	0.0126	0.0021	0.0001
9											1.0000	0.9989	0.9919	0.9682	0.9161	0.8305	0.7166	0.5874	0.4579	0.2424	0.0699	0.0261	0.0050	0.0002
10												0.9997	0.9972	0.9863	0.9574	0.9015	0.8159	0.7060	0.5830	0.3472	0.1185	0.0491	0.0108	0.0006
11												0.9999	0.9991	0.9945	0.9799	0.9467	0.8881	0.8030	0.6968	0.4616	0.1848	0.0847	0.0214	0.0014
12												1.0000	0.9997	0.9980	0.9912	0.9730	0.9362	0.8758	0.7916	0.5760	0.2676	0.1350	0.0390	0.0031
13													0.9999	0.9993	0.9964	0.9872	0.9658	0.9261	0.8645	0.6815	0.3632	0.2009	0.0661	0.0065
14													1.0000	0.9998	0.9986	0.9943	0.9827	0.9585	0.9165	0.7720	0.4657	0.2808	0.1049	0.0124
15														0.9999	0.9995	0.9976	0.9918	0.9780	0.9513	0.8444	0.5681	0.3715	0.1565	0.0223
16														1.0000	0.9998	0.9990	0.9963	0.9889	0.9730	0.8987	0.6641	0.4677	0.2211	0.0377
17															0.9999	0.9996	0.9984	0.9947	0.9857	0.9370	0.7489	0.5640	0.2970	0.0605
18															1.0000	0.9999	0.9993	0.9976	0.9928	0.9626	0.8195	0.6550	0.3814	0.0920
19																1.0000	0.9997	0.9989	0.9965	0.9787	0.8752	0.7363	0.4703	0.1336
20																	0.9999	0.9996	0.9984	0.9884	0.9170	0.8055	0.5591	0.1855
21																	1.0000	0.9998	0.9993	0.9939	0.9469	0.8615	0.6437	0.2473
22																		0.9999	0.9997	0.9970	0.9673	0.9047	0.7206	0.3175
23																		1.0000	0.9999	0.9985	0.9805	0.9367	0.7875	0.3939
24																			1.0000	0.9993	0.9888	0.9594	0.8432	0.4734
25																				0.9997	0.9938	0.9748	0.8878	0.5529
26																				0.9999	0.9967	0.9848	0.9221	0.6294
27																				0.9999	0.9983	0.9912	0.9475	0.7002
28																				1.0000	0.9991	0.9950	0.9657	0.7634
29																					0.9996	0.9973	0.9782	0.8179
30																					0.9998	0.9986	0.9865	0.8633
31																					0.9999	0.9993	0.9919	0.8999
32																					1.0000	0.9996	0.9953	0.9285

Normal estándar										
z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.500	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Normal estándar										
z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998

Student										
g.l.	**0.1000**	**0.0800**	**0.0500**	**0.0250**	**0.0100**	**0.0050**	**0.0025**	**0.0010**	**0.0008**	**0.0005**
1	3.0777	3.8947	6.3138	12.7062	31.8205	63.6567	127.3213	318.3088	397.8865	636.6192
2	1.8856	2.1894	2.9200	4.3027	6.9646	9.9248	14.0890	22.3271	24.9700	31.5991
3	1.6377	1.8589	2.3534	3.1824	4.5407	5.8409	7.4533	10.2145	11.0207	12.9240
4	1.5332	1.7229	2.1318	2.7764	3.7469	4.6041	5.5976	7.1732	7.6104	8.6103
5	1.4759	1.6493	2.0150	2.5706	3.3649	4.0321	4.7733	5.8934	6.1943	6.8688
6	1.4398	1.6033	1.9432	2.4469	3.1427	3.7074	4.3168	5.2076	5.4416	5.9588
7	1.4149	1.5718	1.8946	2.3646	2.9980	3.4995	4.0293	4.7853	4.9805	5.4079
8	1.3968	1.5489	1.8595	2.3060	2.8965	3.3554	3.8325	4.5008	4.6712	5.0413
9	1.3830	1.5315	1.8331	2.2622	2.8214	3.2498	3.6897	4.2968	4.4501	4.7809
10	1.3722	1.5179	1.8125	2.2281	2.7638	3.1693	3.5814	4.1437	4.2845	4.5869
11	1.3634	1.5069	1.7959	2.2010	2.7181	3.1058	3.4966	4.0247	4.1560	4.4370
12	1.3562	1.4979	1.7823	2.1788	2.6810	3.0545	3.4284	3.9296	4.0536	4.3178
13	1.3502	1.4903	1.7709	2.1604	2.6503	3.0123	3.3725	3.8520	3.9700	4.2208
14	1.3450	1.4839	1.7613	2.1448	2.6245	2.9768	3.3257	3.7874	3.9006	4.1405
15	1.3406	1.4784	1.7531	2.1314	2.6025	2.9467	3.2860	3.7328	3.8420	4.0728
16	1.3368	1.4736	1.7459	2.1199	2.5835	2.9208	3.2520	3.6862	3.7919	4.0150
17	1.3334	1.4694	1.7396	2.1098	2.5669	2.8982	3.2224	3.6458	3.7485	3.9651
18	1.3304	1.4656	1.7341	2.1009	2.5524	2.8784	3.1966	3.6105	3.7107	3.9216
19	1.3277	1.4623	1.7291	2.0930	2.5395	2.8609	3.1737	3.5794	3.6774	3.8834
20	1.3253	1.4593	1.7247	2.0860	2.5280	2.8453	3.1534	3.5518	3.6479	3.8495
21	1.3232	1.4567	1.7207	2.0796	2.5176	2.8314	3.1352	3.5272	3.6215	3.8193
22	1.3212	1.4542	1.7171	2.0739	2.5083	2.8188	3.1188	3.5050	3.5978	3.7921
23	1.3195	1.4520	1.7139	2.0687	2.4999	2.8073	3.1040	3.4850	3.5764	3.7676
24	1.3178	1.4500	1.7109	2.0639	2.4922	2.7969	3.0905	3.4668	3.5569	3.7454
25	1.3163	1.4482	1.7081	2.0595	2.4851	2.7874	3.0782	3.4502	3.5392	3.7251
26	1.3150	1.4464	1.7056	2.0555	2.4786	2.7787	3.0669	3.4350	3.5230	3.7066
27	1.3137	1.4449	1.7033	2.0518	2.4727	2.7707	3.0565	3.4210	3.5081	3.6896
28	1.3125	1.4434	1.7011	2.0484	2.4671	2.7633	3.0469	3.4082	3.4943	3.6739
29	1.3114	1.4421	1.6991	2.0452	2.4620	2.7564	3.0380	3.3962	3.4816	3.6594
30	1.3104	1.4408	1.6973	2.0423	2.4573	2.7500	3.0298	3.3852	3.4698	3.6460
31	1.3095	1.4396	1.6955	2.0395	2.4528	2.7440	3.0221	3.3749	3.4588	3.6335
32	1.3086	1.4385	1.6939	2.0369	2.4487	2.7385	3.0149	3.3653	3.4486	3.6218
33	1.3077	1.4375	1.6924	2.0345	2.4448	2.7333	3.0082	3.3563	3.4390	3.6109
34	1.3070	1.4365	1.6909	2.0322	2.4411	2.7284	3.0020	3.3479	3.4300	3.6007
35	1.3062	1.4356	1.6896	2.0301	2.4377	2.7238	2.9960	3.3400	3.4216	3.5911
40	1.3031	1.4317	1.6839	2.0211	2.4233	2.7045	2.9712	3.3069	3.3863	3.5510
50	1.2987	1.4263	1.6759	2.0086	2.4033	2.6778	2.9370	3.2614	3.3378	3.4960
60	1.2958	1.4227	1.6706	2.0003	2.3901	2.6603	2.9146	3.2317	3.3062	3.4602
100	1.2901	1.4156	1.6602	1.9840	2.3642	2.6259	2.8707	3.1737	3.2446	3.3905
120	1.2886	1.4138	1.6577	1.9799	2.3578	2.6174	2.8599	3.1595	3.2295	3.3735
∞	1.2816	1.4051	1.6449	1.9600	2.3263	2.5758	2.8070	3.0902	3.1559	3.2905

Chi-cuadrado

g.l.	0.9975	0.9950	0.9900	0.9750	0.9500	0.0500	0.0250	0.0100	0.0050	0.0025
1	0.0000	0.0000	0.0002	0.0010	0.0039	3.8415	5.0239	6.6349	7.8794	9.1406
2	0.0050	0.0100	0.0201	0.0506	0.1026	5.9915	7.3778	9.2103	10.5966	11.9829
3	0.0449	0.0717	0.1148	0.2158	0.3518	7.8147	9.3484	11.3449	12.8382	14.3203
4	0.1449	0.2070	0.2971	0.4844	0.7107	9.4877	11.1433	13.2767	14.8603	16.4239
5	0.3075	0.4117	0.5543	0.8312	1.1455	11.0705	12.8325	15.0863	16.7496	18.3856
6	0.5266	0.6757	0.8721	1.2373	1.6354	12.5916	14.4494	16.8119	18.5476	20.2494
7	0.7945	0.9893	1.2390	1.6899	2.1673	14.0671	16.0128	18.4753	20.2777	22.0404
8	1.1043	1.3444	1.6465	2.1797	2.7326	15.5073	17.5345	20.0902	21.9550	23.7745
9	1.4501	1.7349	2.0879	2.7004	3.3251	16.9190	19.0228	21.6660	23.5894	25.4625
10	1.8274	2.1559	2.5582	3.2470	3.9403	18.3070	20.4832	23.2093	25.1882	27.1122
11	2.2321	2.6032	3.0535	3.8157	4.5748	19.6751	21.9200	24.7250	26.7568	28.7293
12	2.6612	3.0738	3.5706	4.4038	5.2260	21.0261	23.3367	26.2170	28.2995	30.3185
13	3.1119	3.5650	4.1069	5.0088	5.8919	22.3620	24.7356	27.6882	29.8195	31.8831
14	3.5820	4.0747	4.6604	5.6287	6.5706	23.6848	26.1189	29.1412	31.3193	33.4260
15	4.0697	4.6009	5.2293	6.2621	7.2609	24.9958	27.4884	30.5779	32.8013	34.9496
16	4.5734	5.1422	5.8122	6.9077	7.9616	26.2962	28.8454	31.9999	34.2672	36.4557
17	5.0917	5.6972	6.4078	7.5642	8.6718	27.5871	30.1910	33.4087	35.7185	37.9461
18	5.6233	6.2648	7.0149	8.2307	9.3905	28.8693	31.5264	34.8053	37.1565	39.4221
19	6.1674	6.8440	7.6327	8.9065	10.1170	30.1435	32.8523	36.1909	38.5823	40.8850
20	6.7228	7.4338	8.2604	9.5908	10.8508	31.4104	34.1696	37.5662	39.9968	42.3357
21	7.2889	8.0337	8.8972	10.2829	11.5913	32.6706	35.4789	38.9322	41.4011	43.7751
22	7.8649	8.6427	9.5425	10.9823	12.3380	33.9244	36.7807	40.2894	42.7957	45.2041
23	8.4502	9.2604	10.1957	11.6886	13.0905	35.1725	38.0756	41.6384	44.1813	46.6235
24	9.0442	9.8862	10.8564	12.4012	13.8484	36.4150	39.3641	42.9798	45.5585	48.0337
25	9.6463	10.5197	11.5240	13.1197	14.6114	37.6525	40.6465	44.3141	46.9279	49.4354
26	10.2562	11.1602	12.1981	13.8439	15.3792	38.8851	41.9232	45.6417	48.2899	50.8291
27	10.8733	11.8076	12.8785	14.5734	16.1514	40.1133	43.1945	46.9629	49.6449	52.2153
28	11.4973	12.4613	13.5647	15.3079	16.9279	41.3371	44.4608	48.2782	50.9934	53.5943
29	12.1279	13.1211	14.2565	16.0471	17.7084	42.5570	45.7223	49.5879	52.3356	54.9666
30	12.7646	13.7867	14.9535	16.7908	18.4927	43.7730	46.9792	50.8922	53.6720	56.3325
31	13.4073	14.4578	15.6555	17.5387	19.2806	44.9853	48.2319	52.1914	55.0027	57.6923
32	14.0555	15.1340	16.3622	18.2908	20.0719	46.1943	49.4804	53.4858	56.3281	59.0464
33	14.7092	15.8153	17.0735	19.0467	20.8665	47.3999	50.7251	54.7755	57.6484	60.3949
34	15.3680	16.5013	17.7891	19.8063	21.6643	48.6024	51.9660	56.0609	58.9639	61.7381
35	16.0317	17.1918	18.5089	20.5694	22.4650	49.8018	53.2033	57.3421	60.2748	63.0764
36	16.7001	17.8867	19.2327	21.3359	23.2686	50.9985	54.4373	58.6192	61.5812	64.4098
37	17.3730	18.5858	19.9602	22.1056	24.0749	52.1923	55.6680	59.8925	62.8833	65.7386
38	18.0502	19.2889	20.6914	22.8785	24.8839	53.3835	56.8955	61.1621	64.1814	67.0630
39	18.7317	19.9959	21.4262	23.6543	25.6954	54.5722	58.1201	62.4281	65.4756	68.3831
40	19.4171	20.7065	22.1643	24.4330	26.5093	55.7585	59.3417	63.6907	66.7660	69.6991
41	20.1064	21.4208	22.9056	25.2145	27.3256	56.9424	60.5606	64.9501	68.0527	71.0112
42	20.7994	22.1385	23.6501	25.9987	28.1440	58.1240	61.7768	66.2062	69.3360	72.3195
43	21.4960	22.8595	24.3976	26.7854	28.9647	59.3035	62.9904	67.4593	70.6159	73.6241
44	22.1961	23.5837	25.1480	27.5746	29.7875	60.4809	64.2015	68.7095	71.8926	74.9253
45	22.8996	24.3110	25.9013	28.3662	30.6123	61.6562	65.4102	69.9568	73.1661	76.2229

APÉNDICE C

Resoluciones

En este Apéndice desarrollaremos algunos de los ejercicios propuestos en este libro en forma detallada.

EJERCICIO 1.6

1. Sabemos que al ser A y A^c disjuntos, es decir, $A \cap A^c = \emptyset$, por la Ley aditiva tenemos que:

$$P\big(A \cup A^c|B\big) = P\big(A|B\big) + P\big(A^c|B\big) \tag{C.1}$$

Por otro lado, sabemos que $A \cup A^c = \Omega$ entonces

$$P\big(A \cup A^c|B\big) = P\big(\Omega|B\big) = 1 \tag{C.2}$$

Entonces juntando los resultados de (C.1) y (C.2), es verdadera la afirmación:

$$P\big(A|B\big) + P\big(A^c|B\big) = 1$$

4. Por la definición de probabilidad condicional, si $P(A) > 0$, tenemos que

$$P\big(A|A\big) = \frac{P\big(A \cap A\big)}{P(A)} \tag{C.3}$$

Y como $A \cap A = A$ entonces de (C.3) obtenemos que $P\big(A|A\big) = P(A)/P(A) = 1$. Con lo cual, $P\big(A|A\big) = P(A)$ es falso, pues no vale para todo evento A.

EJERCICIO 2.4

1. Un "ensayo" es extraer una bola y se define como "éxito" que la bola extraída sea blanca y como "fracaso" que la bola extraída sea negra. Si las extracciones se realizan con reemplazo, es decir devolviendo a la urna la bola extraída, se verifica que hay independencia entre las extracciones. Luego la probabilidad de "éxito" (o sea que la bola extraída es blanca) en cada extracción es constante y vale $1/10 = 0.1$. Entonces $X =$ "el número de bolas blancas obtenidas en las 5 extracciones" es una v.a. binomial, $X \sim B(5, 0.1)$ y $v_X = \{0, 1, 2, 3, 4, 5\}$.

2. Observar que el evento "sacar exactamente 2 bolas blancas" es el evento $(X = 2)$, entoces su probabilidad es:

$$P(X = 2) = f(2) = \binom{5}{2} 0.1^2 (1 - 0.1)^{5-2} = 0.0729.$$

3. El evento "obtener al menos 2 bolas blancas" es el evento $(X \geq 2)$, entoces su probabilidad es:

$$\begin{aligned} P(X \geq 2) &= f(2) + f(3) + f(4) + f(5) = 1 - \big[f(0) + f(1)\big] = 1 - f(0) - f(1) \\ &= 1 - \binom{5}{0} 0.1^0 (1 - 0.1)^{5-0} - \binom{5}{1} 0.1^1 (1 - 0.1)^{5-1} = 0.0815. \end{aligned}$$

4. Notar que el evento $(X \leq 3)$, representa al evento "sacar no más de 3 bolas blancas", entonces la probabilidad pedida es:

$$P(X \leq 3) = F(3) = 0.9995 \qquad \big(\text{por Tabla}\big)$$

También está probabilidad se puede calcular como:

$$\begin{aligned} P(X \leq 3) &= 1 - P(X > 3) = 1 - P(X \geq 4) = 1 - [f(4) + f(5)] \\ &= 1 - \binom{5}{4} 0.1^4 (1 - 0.1)^{5-4} - \binom{5}{5} 0.1^5 (1 - 0.1)^{5-5} = 0.9995 \end{aligned}$$

EJERCICIO 2.6

Primero defininimos la v.a. correspondiente para poder resolver este ejercicio: $X_t =$ "número de partículas radiactivas emitidas en t minutos", entonces $X_t \sim P(\lambda_t)$, con $\lambda_t = 6 \times t = 6t$, es decir, $X_t \sim P(6t)$.

1. Para resolver este inciso es claro que $t = 1$, es decir, necesitamos la v.a. $X_1 =$ "número de partículas radiactivas emitidas en un minuto" y $X_1 \sim P(6)$. Luego la probabilidad pedida es:

$$P(X_1 = 0) = f(0) = e^{-6} \frac{6^0}{0!} = e^{-6} = 0.0025$$

2. Sabemos que 30 segundos es igual a medio minuto, es decir que $t = 1/2$, entonces la v.a. que necesitamos aquí es: $X_{1/2}$ = "número de partículas radiactivas emitidas en medio minuto" y $X_{1/2} \sim P(3)$. Luego la probabilidad pedida se calcula como:

$$\begin{aligned} P(X_{1/2} \geq 2) &= 1 - P(X_{1/2} < 2) = 1 - [P(X_{1/2} = 0) + P(X_{1/2} = 1)] \\ &= 1 - [f(0) + f(1)] = 1 - \left[e^{-3}\,\frac{3^0}{0!} + e^{-3}\,\frac{3^1}{1!}\right] = 1 - 4\,e^{-3} = 0.8009 \end{aligned}$$

3. Aquí necesitamos definir dos v.a.: X_2 = "número de partículas radiactivas emitidas entre las 9:10 y las 9:12 AM" e Y_2 = "número de partículas radiactivas emitidas entre las 10:10 y las 10:12 AM". Sabemos que $X_2 \sim P(12)$ (pues $t = 2$) e $Y_2 \sim P(12)$ (también tenemos $t = 2$). Entonces la probabilidad pedida es:

$$\begin{aligned} P(Y_2 = 1 | X_2 = 0) &= \frac{P[(Y_2 = 1) \cap (X_2 = 0)]}{P(X_2 = 0)} && \big(\text{por falta de memoria}\big) \\ &= \frac{P(Y_2 = 1)\,\cancel{P(X_2 = 0)}}{\cancel{P(X_2 = 0)}} = P(Y_2 = 1) \\ &= e^{-12}\,\frac{12^1}{1!} = 12\,e^{-12} = 0.0000737 \end{aligned}$$

4. En este inciso debemos hallar el valor de t para el cual $P(X_t \geq 1) > 0.95$, donde X_t es la v.a. definida al principio de este ejercicio. Luego, si desarrollamos la probabilidad planteada en la desigualdad anterior, tenemos que:

$$\begin{aligned} P(X_t \geq 1) &= 1 - P(X_t < 1) = 1 - P(X_t = 0) = 1 - f(0) = 1 - e^{-6t}\,\frac{(6t)^0}{0!} \\ &= 1 - e^{-6t} \end{aligned}$$

Entonces, ahora resolvamos la desigualdad planteada arriba:

$$\begin{aligned} 1 - e^{-6t} &> 0.95 \\ 0.05 - e^{-6t} &> 0 && \text{(restando de ambos lados 0.95)} \\ 0.05 &> e^{-6t} && \text{(sumando de ambos lados } e^{-6t}) \\ \ln(0.05) &> -6t && \text{(aplicando } \ln \text{ y por propiedad de función inversa)} \\ \frac{\ln(0.05)}{-6} &< t && \text{(dividiendo en ambos lados por } -6) \\ 0.4993 &< t \end{aligned}$$

Por lo tanto, el período de tiempo debe ser mayor a 30 segundos para que la probabilidad de que haya al menos una emisión sea mayor a 0.95.

EJERCICIO 3.3

1. Sabemos que $E(X) = \mu$, $var(X) = \sigma^2$ y $Z = \dfrac{X - \mu}{\sigma}$. Entonces, por un lado:

$$\begin{aligned} E(Z) &= E\left(\frac{X - \mu}{\sigma}\right) = E\left(\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right) && \big(\text{por la Propiedad de Linealidad de la esperanza}\big) \\ &= \frac{1}{\sigma}E(X) - \frac{\mu}{\sigma} = \frac{1}{\cancel{\sigma}}\mu - \frac{\cancel{\mu}}{\cancel{\sigma}} = 0 \end{aligned}$$

Y por otro lado:

$$\begin{aligned} var(Z) &= var\left(\frac{X-\mu}{\sigma}\right) = var\left(\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right) & \big(\text{por (2.10)}\big) \\ &= \left(\frac{1}{\sigma}\right)^2 var(X) = \frac{1}{\sigma^2}\sigma^2 = 1 \end{aligned}$$

EJERCICIO 4.2

Definimos las v.a. X e Y como las medidas del diámetro de la base y altura (en metros) del tanque cilíndrico, respectivamente. Sabemos que $\mu_X = 4.50$, $\sigma_X = 0.10$, $\mu_Y = 1.80$ y $\sigma_Y = 0.06$. Luego el volumen del tanque cilíndrico, a partir de estas dos v.a., es la función:

$$V = h(X,Y) = \pi/2 \times X^2 \times Y$$

Para calcular la desviación típica de V, por la Proposición 4.3 (al ser h un polinomio, es diferenciable en (μ_X, μ_Y)), tenemos que:

$$\begin{aligned} \frac{\partial h}{\partial X}(X,Y) &= \pi/2 \times 2X \times Y = \pi \times X \times Y & \Rightarrow \quad \frac{\partial h}{\partial X}(\mu_X,\mu_Y) = \pi \times 4.50 \times 1.80 = 25.4469 \\ \frac{\partial h}{\partial Y}(X,Y) &= \pi/2 \times X^2 & \Rightarrow \quad \frac{\partial h}{\partial Y}(\mu_X,\mu_Y) = \pi/2 \times 4.50^2 = 31.8086 \end{aligned}$$

Entonces:

$$dt(V) \cong \sqrt{(25.4469)^2 0.10^2 + (31.8086)^2 0.06^2} = \sqrt{6.4754 + 3.6424} = 3.1808$$

es decir, la desviación aproximada del volumen del tanque cilíndrico es de 3.1808 m^2.

EJERCICIO 5.1

2. Sea X el tiempo de vida (en horas) de una lámpara, $X \sim Exp(\lambda)$, entonces por la Proposición 3.3, $E(X) = 1/\lambda$. Por un lado, como la función $h(\lambda) = 1/\lambda$ es continua para $\lambda > 0$ tenemos que:

$$\widehat{h(\lambda)} = h(\widehat{\lambda}) = 1/\widehat{\lambda} \tag{C.4}$$

Por otro lado, como mencionamos en el Capítulo 5, el estimador usual para la media (o esperanza), es la media muestral $\overline{X}$, es decir

$$\widehat{E(X)} = \overline{X} \tag{C.5}$$

Entonces igualando (C.4) y (C.5) obtenemos que el estimador para el parámetro λ es:

$$\widehat{\lambda} = 1/\overline{X} \tag{C.6}$$

Ahora veamos cual es su estimación: se los datos tenemos que $\bar{x} = 22.79$ entonces, reemplazando este valor en (C.6), finalmente obtenemos $\widehat{\lambda} = 0.0439$.

Para obtener ahora la estimación de la probabilidad pedida, primero planteamos esta probabilidad que deseamos estimar:

$$P(X > 50) = 1 - P(X \leq 50) = 1 - F(50) = 1 - \left(1 - e^{-\lambda 50}\right) = e^{-\lambda 50}$$

Como vemos que esta probabilidad es una función continua del parámetro, tenemos que

$$P(\widehat{X > 50}) = e^{-\widehat{\lambda} 50} = e^{-0.0439 \times 50} = 0.1114$$

es la probabilidad estimada de que una lámpara de ese tipo dure más de 50 horas.

EJERCICIO 7.2

1. En el Ejemplo 5.6, se demostró que dada una m.a. com media μ y varianza σ^2, la varianza muestral, S^2, es un estimador insesgado de σ^2, es decir, $E(S^2) = \sigma^2$. Entonces, tenemos que, si S_1^2 y S_2^2 son las varianzas muestrales de $X_1, X_2, \ldots, X_{n_1}$ y $Y_1, Y_2, \ldots, Y_{n_2}$ respectivamente, luego:

$$E(S_1^2) = \sigma^2 = E(S_2^2) \tag{C.7}$$

pues ambas m.a. tienen la misma varianza. Ahora, calculemos la esperanza de S_p^2:

$$\begin{aligned} E(S_p^2) &= E\left(\frac{(n_1-1)\, S_1^2 + (n_2-1)\, S_2^2}{n_1+n_2-2}\right) \\ &= \frac{(n_1-1)\, E(S_1^2) + (n_2-1)\, E(S_2^2)}{n_1+n_2-2} && \big(\text{por la Propiedad de Linealidad de la esperanza}\big) \\ &= \frac{(n_1-1)\, \sigma^2 + (n_2-1)\, \sigma^2}{n_1+n_2-2} && \big(\text{por (C.7)}\big) \\ &= \frac{\sigma^2\, \cancel{(n_1+n_2-2)}}{\cancel{n_1+n_2-2}} = \sigma^2 && \big(\text{sacando factor común } \sigma^2\big) \end{aligned}$$

Por lo tanto, S_p^2 es un estimador insesgado de σ^2.

EJERCICIO 8.1

1. Comencemos probando la igual de S_{xy}:

$$\begin{aligned}
S_{xy} &= \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n}\left(x_iy_i - \bar{x}y_i - \bar{y}x_i + \bar{x}\,\bar{y}\right) && \text{(distribuyendo)}\\
&= \sum_{i=1}^{n} x_iy_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + n\,\bar{x}\,\bar{y} && \text{(distribuyendo)}\\
&= \sum_{i=1}^{n} x_iy_i - n\,\bar{x}\,\bar{y} - \cancel{n\,\bar{y}\,\bar{x}} + \cancel{n\,\bar{x}\,\bar{y}} && \left(\text{pues} \sum_{i=1}^{n} x_i = n\bar{x} \ \text{ y } \ \sum_{i=1}^{n} y_i = n\bar{y}\right)\\
&= \sum_{i=1}^{n} x_iy_i - n\,\bar{x}\,\bar{y}
\end{aligned}$$

Siguiendo los mismos pasos se llega fácilmente a $S_{xx} = \sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2$ y $S_{yy} = \sum_{i=1}^{n} y_i^2 - n\,\bar{y}^2$.

2. Veamos que $S_{rr} = \sum_{i=1}^{n}\left(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)\right)^2 = \sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2$ se mimimiza cuando $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x}$ y $\widehat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$. Realicemos las derivadas parciales:

$$\frac{\partial S_{rr}}{\partial \widehat{\beta}_0} = \sum_{i=1}^{n}\left(2(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)(-1)\right) = -2\left(\sum_{i=1}^{n} y_i - n\widehat{\beta}_0 - \widehat{\beta}_1\sum_{i=1}^{n} x_i\right) \tag{C.8}$$

$$\frac{\partial S_{rr}}{\partial \widehat{\beta}_1} = \sum_{i=1}^{n}\left(2(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)(-x_i)\right) = -2\left(\sum_{i=1}^{n} y_ix_i - \widehat{\beta}_0\sum_{i=1}^{n} x_i - \widehat{\beta}_1\sum_{i=1}^{n} x_i^2\right) \tag{C.9}$$

Igualando las derivadas parciales a cero, en la ecuación (C.8) tenemos que:

$$\sum_{i=1}^{n} y_i - n\widehat{\beta}_0 - \widehat{\beta}_1\sum_{i=1}^{n} x_i = n\bar{y} - n\widehat{\beta}_0 - \widehat{\beta}_1 n\bar{x} = 0$$

o equivalentemente:

$$\bar{y} - \widehat{\beta}_0 - \widehat{\beta}_1\bar{x} = 0$$

y despejando obtenemos que

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x} \tag{C.10}$$

En la ecuación (C.9) tenemos que:

$$\sum_{i=1}^{n} y_ix_i - \widehat{\beta}_0\sum_{i=1}^{n} x_i - \widehat{\beta}_1\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_ix_i - \widehat{\beta}_0 n\bar{x} - \widehat{\beta}_1\sum_{i=1}^{n} x_i^2 = 0 \tag{C.11}$$

y reemplazando (C.10) en (C.11) obtenemos:

$$\sum_{i=1}^{n} y_ix_i - (\bar{y} - \widehat{\beta}_1\bar{x})n\bar{x} - \widehat{\beta}_1\sum_{i=1}^{n} x_i^2 = 0$$

o equivalentemente, sacando factor común $\widehat{\beta}_1$ y utilizando las igualdades de S_{xy} y S_{xx} antes demostradas, tenemos:

$$\sum_{i=1}^{n} y_ix_i - \bar{y}n\bar{x} - \widehat{\beta}_1\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = S_{xy} - \widehat{\beta}_1 S_{xx} = 0$$

y despejando, finalmente obtenemos que $\widehat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$.

EJERCICIO 8.2

Para el modelo de regresión lineal simple $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$ bajo las siguientes hipótesis:

- $E(\epsilon_i) = 0$, $i = 1, \ldots, n$,
- $var(\epsilon_i) = \sigma^2$, $i = 1, \ldots, n$, y
- ϵ_i y ϵ_j son independientes entre sí, con $i \neq j$,

se obtiene que, las v.a. Y_i son independientes entre sí y:

$$E(Y_i) = E(\beta_0 + \beta_1 x_i + \epsilon_i) \overset{(1)}{=} \beta_0 + \beta_1 x_i + E(\epsilon_i) = \beta_0 + \beta_1 x_i \tag{C.12}$$

$$E(\overline{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \overset{(1)}{=} \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) \overset{(2)}{=} \beta_0 + \beta_1 \overline{x} \tag{C.13}$$

$$var(Y_i) = var(\beta_0 + \beta_1 x_i + \epsilon_i) \overset{(3)}{=} var(\epsilon_i) = \sigma^2 \tag{C.14}$$

donde en (1) se utilizó la Propiedad de Linealidad de la esperanza, en (2) se utilizaron las igualdades $\sum_{i=1}^{n} \beta_0 = n\beta_0$ y $\sum_{i=1}^{n} x_i = n\overline{x}$ y en (3) es por (2.10)

Comencemos probando que $\widehat{\beta}_1$ es un estimador insesgados de β_1:

$$\begin{aligned}
E(\widehat{\beta}_1) &= E\left(\frac{S_{xy}}{S_{xx}}\right) = E\left(\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})\right) && \big(\text{definición de } S_{xy}\big)\\
&= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \overline{x})\big(E(Y_i) - E(\overline{Y})\big) && \big(\text{Prop. de linealidad de la esperanza}\big)\\
&= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \overline{x})\Big(\cancel{\beta_0} + \beta_1 x_i - (\cancel{\beta_0} + \beta_1 \overline{x})\Big) && \big((\text{C.12}) \text{ y } (\text{C.13})\big)\\
&= \frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i - \overline{x})\beta_1(x_i - \overline{x}) = \frac{\beta_1}{S_{xx}}\sum_{i=1}^{n}(x_i - \overline{x})^2 && \big(\text{factor común } \beta_1\big)\\
&= \frac{\beta_1}{\cancel{S_{xx}}}\cancel{S_{xx}} = \beta_1 && \big(\text{definición de } S_{xx}\big)
\end{aligned}$$

Ahora demostramos que $\widehat{\beta}_0$ es un estimador insesgados de β_0:

$$\begin{aligned}
E(\widehat{\beta}_0) &= E\left(\overline{Y} - \widehat{\beta}_1 \overline{x}\right) = E(\overline{Y}) - E(\widehat{\beta}_1)\,\overline{x} && \big(\text{por Prop. de linealidad de la esperanza}\big)\\
&= (\beta_0 + \cancel{\beta_1 \overline{x}}) - \cancel{\beta_1 \overline{x}} = \beta_0 && \big(\text{por (C.13) y por ser } \widehat{\beta}_1 \text{ insesgado}\big)
\end{aligned}$$

Así queda demostrado (8.3)

Para llegar a la expresión de la varianza de $\widehat{\beta}_0$ necesitamos expresar a este estimador de otra forma equivalente, como combinación lineal de las Y_i:

$$\begin{aligned}
\widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{x} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \overline{x}\,\frac{S_{xy}}{S_{xx}} && \big(\text{definiciones de } \overline{Y} \text{ y } \widehat{\beta}_1\big)\\
&= \sum_{i=1}^{n}\frac{1}{n}\,Y_i - \frac{\overline{x}}{S_{xx}}\,\sum_{i=1}^{n}(x_i - \overline{x})Y_i && \big(\text{expresión } S_{xy}\big)\\
&= \sum_{i=1}^{n}\left(\frac{1}{n}Y_i - \frac{\overline{x}}{S_{xx}}(x_i - \overline{x})Y_i\right) && \big(\text{juntando en una sumatoria}\big)\\
&= \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\overline{x}(x_i - \overline{x})}{S_{xx}}\right)Y_i && \big(\text{factor común } Y_i\big)
\end{aligned}$$

Ahora sí calculemos la varianza de $\widehat{\beta}_0$:

$$
\begin{aligned}
var(\widehat{\beta}_0) &= var\left[\sum_{i=1}^{n}\left(\frac{1}{n}-\frac{\overline{x}(x_i-\overline{x})}{S_{xx}}\right)Y_i\right] \\
&= \sum_{i=1}^{n}\left(\frac{1}{n}-\frac{\overline{x}(x_i-\overline{x})}{S_{xx}}\right)^2 var(Y_i) = \sum_{i=1}^{n}\left(\frac{1}{n}-\frac{\overline{x}(x_i-\overline{x})}{S_{xx}}\right)^2\sigma^2 && \big((2.10)\text{ y (C.14)}\big) \\
&= \sum_{i=1}^{n}\left(\frac{1}{n^2}-2\frac{1}{n}\frac{\overline{x}(x_i-\overline{x})}{S_{xx}}+\frac{\overline{x}^2(x_i-\overline{x})^2}{S_{xx}^2}\right)\sigma^2 && \big(\text{resolviendo el cuadrado}\big) \\
&= \sigma^2\left(\sum_{i=1}^{n}\frac{1}{n^2}-\sum_{i=1}^{n}2\frac{1}{n}\frac{\overline{x}(x_i-\overline{x})}{S_{xx}}+\sum_{i=1}^{n}\frac{\overline{x}^2(x_i-\overline{x})^2}{S_{xx}^2}\right) \\
&= \sigma^2\left(\cancel{n}\frac{1}{n^{\cancel{2}}}-2\frac{1}{n}\frac{\overline{x}}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})+\frac{\overline{x}^2}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\overline{x})^2\right) \\
&= \sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}^{\cancel{2}}}\cancel{S_{xx}}\right) && \left(\sum_{i=1}^{n}(x_i-\overline{x})=0\right) \\
&= \sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right)
\end{aligned}
$$

Finalmente calculemos la varianza de $\widehat{\beta}_1$:

$$
\begin{aligned}
var(\widehat{\beta_1}) &= var\left(\frac{S_{xy}}{S_{xx}}\right) = var\left(\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})(Y_i-\overline{Y})\right) && \big(\text{por definición de } S_{xy}\big) \\
&= var\left(\frac{1}{S_{xx}}\sum_{i=1}^{n}(x_i-\overline{x})Y_i\right) = \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\overline{x})^2 var(Y_i) && \big(\text{por (*) y por (2.10)}\big) \\
&= \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\overline{x})^2\sigma^2 = \frac{\sigma^2}{S_{xx}^2}\sum_{i=1}^{n}(x_i-\overline{x})^2 && \big(\text{por (C.14)}\big) \\
&= \frac{\sigma^2}{S_{xx}^{\cancel{2}}}\cancel{S_{xx}} = \frac{\sigma^2}{S_{xx}} && \big(\text{por definición de} S_{xx}\big)
\end{aligned}
$$

donde en (*) $\sum_{i=1}^{n}(x_i-\overline{x})(Y_i-\overline{Y}) = \sum_{i=1}^{n}(x_i-\overline{x})Y_i - \overline{Y}\sum_{i=1}^{n}(x_i-\overline{x}) = \sum_{i=1}^{n}(x_i-\overline{x})Y_i$.

Por lo tanto, queda demostrado (8.4).

Los autores

Coordinadoras

Apezteguía, María Carmen

Licenciada en Matemática Aplicada, Orientación en Investigación Operativa y Estadística de la Facultad de Ciencias Exactas, Universidad Nacional de La Plata. Profesora Asociada con Semi Dedicación del área Probabilidades y Estadística de la Facultad de Ciencias Exactas, Universidad Nacional de La Plata, siendo Coordinadora de la Cátedra de Análisis de Datos. Profesora Titular de la Facultad de Ciencias Económicas, Universidad Nacional de La Plata, en la Cátedra de Estadística para los Negocios desde el año 2015. Profesional Principal de la Carrera del Personal de Apoyo de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires desde el año 1985 hasta el año 2013.

Ferrario, Julieta

Licenciada en Matemática y Doctora de la Facultad de Ciencias Exactas, área Matemática, Universidad Nacional de La Plata. Jefa de Trabajos Prácticos de la Facultad de Ciencias Económicas, Universidad Nacional de La Plata, en la Cátedra Estadística II desde el año 2017. Ayudante Diplomada con Dedicación Exclusiva de la Facultad de Ciencias Exactas, Universidad Nacional de La Plata, en la Cátedra Análisis de Datos desde el año 2015. Directora de la Dirección de Estadísticas de la Facultad de Ciencias Exactas, Universidad Nacional de La Plata, desde el año 2016. Docente

Investigador Categoría V en el área de Estadística. Becaria del Consejo Nacional de Investigaciones Científicas y Técnicas desde el año 2007 hasta el año 2012.

Autoras

D' Urzo, Paula Gisela

Profesora de Matemática y Especialista en Educación de Ciencias Exactas y Naturales de la Facultad de Humanidades y Ciencias de la Educación, Universidad Nacional de La Plata. Profesora Adjunta y Jefa de Trabajos Prácticos de la Facultad de Ingeniería, Universidad Nacional de La Plata, en la cátedra Probabilidades y Estadística desde el año 2017. Ayudante Diplomada de la Facultad de Ciencias Exactas, Universidad Nacional de La Plata, en la cátedra Análisis de Datos desde el año 2010. Ayudante Diplomada de la Facultad de Ciencias Económicas, Universidad Nacional de La Plata, en la Cátedra Estadística II desde el año 2015.

Fasano, Silvia Elena

Licenciada en Matemática Aplicada, Orientación en Investigación Operativa y Estadística de la Facultad de Ciencias Exactas, Universidad Nacional de La Plata. Profesora Adjunta del área de Probabilidades y Estadística la Facultad de Ciencias Exactas de la Universidad Nacional de La Plata, la Cátedra de Análisis de Datos. Profesional Principal de la Carrera del Personal de Apoyo de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, desde el año 1981.

Lavié, Julieta Leonor

Alumna de la carrera Licenciatura en Matemática de la Facultad de Ciencias Exactas, Universidad Nacional de La Plata. Ayudante Alumna de la Facultad de Ciencias Exactas, Universidad Nacional de La Plata, en la Cátedra Análisis de Datos desde el año 2013. Ayudante Alumna de la Facultad de Ingeniería, Universidad Nacional de La Plata, en la cátedra Probabilidades y Estadística desde el año 2015. Directora de la Dirección de Encuestas de la Facultad de Ciencias Exactas, Universidad Nacional de La Plata desde el año 2017.

Suárez, Virginia Elizabeth

Profesora de Matemática de la Facultad de Humanidades y Ciencias de la Educación, Universidad Nacional de La Plata. Jefa de Trabajos Prácticos de la Facultad de Ciencias Exactas, Universidad Nacional de La Plata, en la Cátedra Análisis de Datos desde el año 2015. Ayudante Diplomada de la Facultad de Ciencias Exactas, Universidad Nacional de La Plata, en la Cátedra Matemática II para alumnos de la Facultad Informática, Universidad Nacional de La Plata, desde el año 2016. Ayudante Diplomada de la Facultad de Ingeniería, Universidad Nacional de La Plata, en la cátedra Probabilidades y Estadística desde el año 2015. Ayudante Diplomada de la Facultad de Ciencias Económicas, Universidad Nacional de La Plata, en la Cátedra de Estadística para los Negocios desde el año 2017.

Libros de Cátedra

Probabilidades y Estadística : análisis de datos / María Carmen Apezteguía ... [et al.] ; coordinación general de María Carmen Apezteguía ; Julieta Ferrario. - 1a edición para el alumno - La Plata : Universidad Nacional de La Plata ; La Plata : EDULP, 2019. Libro digital, PDF - (Libros de cátedra)

Archivo Digital: descarga
ISBN 978-950-34-1735-5

1. Probabilidades. 2. Estadísticas. I. Apezteguía, María Carmen. II. Apezteguía, María Carmen, coord. III. Ferrario, Julieta, coord.
CDD 519.2

Diseño de tapa: Dirección de Comunicación Visual de la UNLP

Universidad Nacional de La Plata – Editorial de la Universidad de La Plata
47 N.º 380 / La Plata B1900AJP / Buenos Aires, Argentina
+54 221 427 3992 / 427 4898
edulp.editorial@gmail.com
www.editorial.unlp.edu.ar

Edulp integra la Red de Editoriales Universitarias Nacionales (REUN)

Primera edición, 2018 ISBN
ISBN 978-950-34-1735-5





